Machine Learning-based Analysis of the Relationship Between the Human Gut Microbiome and Bone Health

by

Pravallika Reddy Ketha Hazarath

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved October 2020 by the
Graduate Supervisory Committee:

Daniel Bliss, Chair
Corrie Whisner
Gautam Dasarathy

ARIZONA STATE UNIVERSITY

December 2020

ABSTRACT

The Human Gut Microbiome (GM) modulates a variety of structural, metabolic, and protective functions to benefit the host. A few recent studies also support the role of the gut microbiome in the regulation of bone health.

The relationship between GM and bone health was analyzed based on the data collected from a group of twenty-three adolescent boys and girls who participated in a controlled feeding study, during which two different doses(0 g/d fiber and 12 g/d fiber) of Soluble Corn Fiber (SCF) were added to their diet.

This analysis was performed by predicting measures of Bone Mineral Density (BMD) and Bone Mineral Content (BMC) which are indicators of bone strength, using the GM sequence of proportions of 178 microbes collected from 23 subjects, by building a machine learning regression model.

The model developed was evaluated by calculating performance metrics such as Root Mean Squared Error, Pearson's correlation coefficient, and Spearman's rank correlation coefficient, using cross-validation.

A noticeable correlation was observed between the GM and bone health, and it was observed that the overall prediction correlation was higher with SCF intervention ($r \approx 0.51$). The genera of microbes that played an important role in this relationship were identified. *Eubacterium (g)*, *Bacteroides (g)*, *Megamonas (g)*, *Acetivibrio (g)*, *Faecalibacterium (g)*, and *Paraprevotella (g)* were some of the microbes that showed an increase in proportion with SCF intervention.

ACKNOWLEDGMENTS

I would like to thank my advisor and committee chair, Professor Daniel Bliss, for giving me the wonderful opportunity to work on this project and for providing invaluable guidance and motivation through the course of this research.

I would like to express my sincere appreciation to Professor Gautam Dasarathy and Professor Corrie Whisner, for willing to serve as members of the defense committee and for providing their insightful suggestions and feedback.

I would also like to thank Arindam Dutta and Owen Ma for their constant help, valuable advice, and for offering assistance with the theoretical and practical aspects of this project.

I am extremely grateful to my family and friends for their unconditional support and continuous encouragement throughout this journey.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Motivation

The Human Gut Microbiome (GM) benefits the host by modulating a wide variety of structural, metabolic, and protective functions. In the community of health science, the complicated interactions between the GM and characteristics of the host and the subsequent functional benefits are an area of great interest and are currently being explored (Wallace *et al.* (2017)). One of the major and upcoming fields being inspected is the role of GM in the regulation of bone health.

Previous studies have indicated that the GM plays an important role in influencing bone physiology by modulating the processes of bone gain and bone loss due to changes in bacterial composition. The GM is also known to modulate bone morphology which in turn affects the bone strength and consequently fracture risk (Medina-Gomez (2018)). Further investigation in the field of understanding the relationship between the GM and bone strength is necessary and is a promising area of research in improving bone health and reducing bone diseases such as osteoporosis (Chen *et al.* (2017)).

However, current research studies are limited in number and are mainly focused towards using animal models. Therefore, more studies focusing on the human gut microbiome are needed. In the majority of previous studies, it has been shown that prebiotic intervention in diet has the ability to induce changes in the gut microbiome and is associated with increased calcium absorption in animal models and humans (Weaver (2015), Whisner and Castillo (2018)). However the number of studies eval-

uating the effect of prebiotics on bone health indices such as bone mineral density and bone mineral content and understanding the potential gut-bone relationships are scarce (Whisner and Castillo (2018)).

In this work, the influence of the gut microbiota on bone health due to the intervention of a prebiotic fiber called Soluble Corn Fiber (SCF) in the diet is studied in human subjects.

1.2 Objective

The main objective of this thesis was to analyze the relationship between the human gut microbiome (GM) and bone health in adolescents, and establish a relationship between them based on the influence of addition of different doses of Soluble Corn Fiber (SCF) to the diet, using Machine Learning techniques. This analysis was performed by predicting measures of Bone Mineral Density (BMD) and Bone Mineral Content (BMC) which are indicators of bone strength, using the gut microbiome sequences collected from subjects participating in a two-phase controlled feeding study, by building regression models.

The goal was to develop a regression model to understand the relationship between the microbes in the gut and the measures of bone health. The regression model was optimized by tuning the model hyperparameters using cross-validation. The performance of the model developed was evaluated by calculating performance metrics such as Root Mean Squared Error, Pearson's correlation coefficient, and Spearman's rank correlation coefficient, using cross validation.

Feature Selection techniques were implemented on the regression model, to identify the genera of microbes that played an important role in affecting the gut-bone relationship, using cross-validation. The effect of SCF was studied on the most relevant microbes by comparing the results from both phases of this experiment.

Current research work indicates that the gut microbiome plays an important role in bone metabolism and the effect of the gut microbiota on bone has been studied mainly in animal models. In humans, a prebiotic intervention resulted in greater whole body bone accrual in adolescents, but this link to changes in the gut microbiome was not studied. Therefore, this project aims to study the effect of the prebiotic fiber (SCF) on bone health (BMD and BMC) and the composition of the gut microbiota.

1.3 Thesis Organization

In Chapter 2, a background study on the human microbiome and gut microbiome is provided. An insight into the previous work conducted in the field of gut microbiome and its influence on bone health is provided. In Chapter 3, a background on the different machine learning techniques used in previous microbiome studies is given and the motivation to use specific algorithms, in this project, is discussed. In Chapter 4, the experiment in focus is discussed along with the methods used to collect the data, and details about the entire dataset are mentioned. An overview of the problem, the different Machine Learning concepts that were used along the course of this project and the procedure followed to perform prediction analysis is highlighted in Chapter 5. All the results and observations obtained during this project are explained in Chapter 6. Finally, Chapter 7 concludes this research with a summary of the contributions made by this research and discusses some possible future research work.

Chapter 2

BACKGROUND ON THE GUT MICROBIOME

2.1 The Gut Microbiota

The human microbiome consists of all microbiota that are present on or within human tissues and biofluids along with the respective locations in which they thrive such as the lungs, skin, saliva, placenta, seminal fluid, uterus, and gastrointestinal tract. The different types of human microbiota include bacteria, archaea, fungi, protists and viruses. The microorganisms are referred to as "microbiota" and the organisms along with their genetic compositions, collectively form the "microbiome" (Yang *et al.* (2016)).

The gut microbiota comprises a variety of microorganisms that live in the digestive tracts of humans and other animals including insects. The microorganisms in the gut exhibit a greater diversity and are more abundant in comparison to other body sites such as the skin, oral cavity and the urogenital tract (Yang *et al.* (2016)). Approximately 300 to 500 bacterial species, consisting of around 2 million genes, exist within the human gastrointestinal microbiota (Quigley (2013)). The gut flora is established at one to two years after birth and its composition changes over time, when the diet changes, and as overall health changes. This composition is different in different parts of the digestive tract. *Firmicutes, Bacteroidetes, Actinobacteria,* and *Proteobacteria* are generally identified as the most dominant phyla.

2.2 Effect of the Gut Microbiota on Human Health

The relationship between some gut flora and humans is a mutualistic relationship. The gut microbiota regulate a number of essential functions associated with digestion of food, activation of the immune system, and control of cognitive processes (Yang *et al.* (2016)). A normal physiology is maintained in the host as long as there is symbiosis of the gut microbiota, whereas a dysbiosis of the microbiota affects the balance and may lead to a variety of diseases (Wang *et al.* (2016)).

- Metabolism: The gut bacteria play an important role in synthesizing amino acids and vitamins such as Vitamin B and Vitamin K, and metabolizing bile acids (Bull and Plummer (2014)). Intestinal bacteria also digest complex carbohydrates and proteins present in the diet, that are not absorbed in the upper gut, by producing hydrolytic enzymes (Macfarlane and Macfarlane (2012)). Short-chain fatty acids (SCFAs), such as acetic acid and butyric acid, are the major products of bacterial fermentation of dietary fiber in the gut. In this process the host is benefitted by absorbing these substrates and retrieving energy, along with the bacteria receiving a supply of energy and nutrients for their growth and survival (Bull and Plummer (2014)).

- Development of Immune System: Most of the bacteria do not allow pathogens to thrive in the intestine by producing antimicrobial substances called bacteriocins and competing for nutrients and sites of attachment in the gut lining (Bull and Plummer (2014)). The gut microbiome controls the body's response to infection by communicating with immune cells.

- Gut–Brain Axis: The gut–brain axis is a bidirectional communication system enabling the intestinal microbiota to access the brain and vice versa. A route is es-

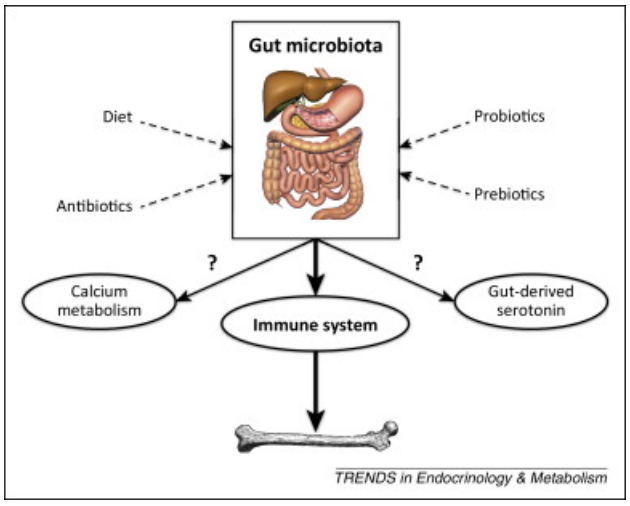


Figure 2.1: Human Gut Mirobiota and Bone Health. From Ohlsson and Sjögren (2015)

tablished, through which the brain can control functions such as peristalsis and mucin production along with other immunity based functions (Bull and Plummer (2014)).

- Gut Microbiota and Disease: Perturbations in the microbial composition can lead to a wide variety of inflammatory conditions affecting the inner and outer parts of the gut. These include inflammatory bowel diseases, rheumatoid arthritis, multiple sclerosis, and asthma, as well as metabolic diseases, such as diabetes and obesity (Ohlsson and Sjögren (2015), Bull and Plummer (2014)).

2.3 Gut Microbiota and Bone Health

Previous studies have shown that the gut microbiome is involved in the regulation of a plethora of biological processes such as gut physiology, nutrient production and absorption, host growth, energy balancing, metabolic functions, immune-system functions, brain– behavior systems, and inflammatory processes (Chen *et al.* (2017)).

Intestinal microbiota also plays an important role in monitoring the health of locations that are away from the intestine including the skin, lungs, arteries, and bone

(McCabe *et al.* (2015)). Although researchers have only recently begun to study the relationship between the gut microbiota and bone metabolism, there are many studies supporting the role of gut microbiome in the regulation of bone density and health.

The GM affects bone metabolism and bone mass by altering the immune system of the host, as there is a well established connection between the immune system and bone metabolism, which was observed in germ-free mice. In a certain experiment by (Sjögren *et al.* (2012)), it was seen that the germ-free mice showed an increase in BMD in comparison to conventionally raised mice. There was a reduction in bone mass when the germ-free mice were colonized with a normal gut microbiota which indicated that the absence of the gut microbiota might be responsible for the increased BMD in the germ-free mice (Chen *et al.* (2017)).

Several studies have indicated that the intestinal microbiota regulate bone mass through different mechanisms such as mediating the immune system, releasing neurotransmitters, and calcium absorption in the intestine (Wang *et al.* (2017)). Previous studies also indicate that the utilization of prebiotics, probiotics, or antibiotic treatment affect the composition of the gut microbiome and in-turn leads to regulation of bone metabolism (Ohlsson and Sjögren (2015)), as seen in Figure 2.1.

2.4 Influence of Gut Microbiome on Bone Health: Effect of Prebiotics

Approaches such as direct modulation of the quantity of bacteria present in the gut through use of antibiotics, addition of bacterial substrates known as prebiotics, and addition of beneficial bacteria called probiotics, to the diet can increase calcium absorption and enhance bone properties. This strategy leads to an overall improvement in bone health (Wallace *et al.* (2017), McCabe *et al.* (2015)).

Analysis indicates that the majority of the population, especially adolescents, consume diets with fewer vegetables, whole grains, and fruits (Krebs-Smith *et al.* (2010)).

There is also an under-consumption of milk and milk products with respect to the recommended quantities (Krebs-Smith *et al.* (2010)). This leads to a deficiency in calcium, vitamin D, and potassium nutrient intakes (Fulgoni III *et al.* (2011)), and further increases the risk of reduced peak bone mass development and eventually increases the risk of osteoporosis during adulthood (Whisner *et al.* (2016)).

Adolescence is the most important time for building a strong skeleton and is the most critical period for bone mineral accrual (Loud and Gordon (2006)). During this period, the hormones of puberty speed up growth of bones and increase bone size and strength. An optimal environment must be established during the years of growth and maturation for the achievement of peak bone mass in order to maintain a strong lifelong bone health and prevent adult osteoporosis and bone fractures (Carey and Golden (2015), Levine (2012)).

For the growth and development of healthy bones and bone mineralization it is required that adolescents receive adequate calcium and vitamin D through a well-balanced nutrition and participate in regular physical activities (Carey and Golden (2015), Cheng *et al.* (2020), Bailey *et al.* (2010)). Long-lasting improvements in bone mineral density might occur with an increase in calcium intake from dairy sources (Levine (2012)).

Despite much public health awareness of the importance of calcium intake for osteoporosis prevention, many do not choose calcium-rich foods (Weaver (2015)). Therefore it is essential to develop an alternative strategy for improving calcium nutrition. One method is to enhance the absorption of any calcium present in the diet with the help of prebiotic dietary fibers, such as nondigestible oligosaccharides and polysaccharides (Whisner *et al.* (2016),Whisner *et al.* (2014)).

The gut microbiota is very plastic, and its composition can be altered rapidly when there is a change in diet (Clarke *et al.* (2014)). The intake of carbohydrates and other

nutrients provide an energy source for the gut bacteria to thrive (Chen *et al.* (2017)). Significant changes in bacterial metabolism associated with small chain fatty acids and amino acids was seen with dietary changes in mice in a short span of one week (Ursell *et al.* (2012)). Studies showed that switching from a low-fat, polysaccharide-rich diet to a high-fat, high-sugar diet led to a dramatic shift in the structure of the human microbiota within a single day (Clarke *et al.* (2014), David *et al.* (2014)) and led to a rapid change in the configuration of the microbiota of humanized gnotobiotic mice (Turnbaugh *et al.* (2009)).

Dietary fiber positively shapes the composition of the gut microbiota and immunity (Shen and Wong (2016)) and fermentation of fiber by gut bacteria has numerous potential health benefits (Klosterbuer *et al.* (2013)). It is hypothesized that the fermentation of soluble fibers in the colon, by intestinal microflora, leads to the production of Short Chain Fatty Acids(SCFAs) (Klosterbuer *et al.* (2013), Zafar *et al.* (2004)). The most abundant SCFAs formed are acetate, propionate and butyrate, with butyrate being considered as the most important for colonic health due to its effects on promoting normal colonocyte development (Klosterbuer *et al.* (2013)). There is a reduction in the luminal pH by these organic acids and this leads to conversion of any unabsorbed calcium which comes from the upper part of the intestine into the ionic form. The SCFAs and low pH medium cause the surface area of the intestine to enlarge and thereby enhance calcium absorption and subsequent utilization (Zafar *et al.* (2004), Weaver *et al.* (2010)). Thus, dietary fibers benefit bone health by undergoing fermentation in the gut and increasing mineral absorption and retention (Weaver *et al.* (2010)).

Maathuis *et al.* (2009) define Prebiotics as "a non-digestible food ingredient that beneficially affects the host by selectively stimulating the growth and/or activity of one or a limited number of bacteria in the colon". Prebiotic fibers by definition are resistant to absorption and are fermented by microbial flora in the large intestine (Weaver

(2015), Scholz-Ahrens *et al.* (2002)).

Adding bacterial fermentative substrates such as complex carbohydrates leads to an increase in concentration of SCFAs, including butyrate, and affects the immune function directly (Neish (2009)). Non-digestible oligosaccharides (NDO) such as inulin, oligofructose and fructo-oligosaccharides, and galacto-oligosaccharides, along with resistant starch, lactulose, and maltitol have also been shown to promote absorption of various minerals such as calcium, magnesium, and zinc. This process has an important impact on the regulation of Bone Mineral Density and the prevention of bone loss and osteoporosis (Chen *et al.* (2017), Scholz-Ahrens *et al.* (2002)).

A series of studies have shown the effects of dietary prebiotics on microbiota, calcium absorption, and bone measures in rats and humans. It was seen that feeding healthy growing rats with 5% fructo-oligosaccharides (Lobo *et al.* (2006)) and supplementing diets with dietary galacto-oligosaccharides (Weaver *et al.* (2011)) led to an increase in the intestinal absorption of calcium and magnesium. This led to an increase in bone mineralization, which, in turn, improved the resistance to fracture (Lobo *et al.* (2006)). An increase in calcium and magnesium retention, bone strength, and bone mineral density was also observed (Weaver *et al.* (2011)).

In other studies related to ovariectomized rats, the incorporation of the following in the diet, fructo-oligosaccharides (Devareddy *et al.* (2006)), Nondigestible Oligosaccharides (Zafar *et al.* (2004)), and oligofructose (Scholz-Ahrens *et al.* (2002)), improved calcium absorption and retention (Zafar *et al.* (2004)) and significantly increased bone mineral density of the lumbar vertebrae, tibiae, whole body (Devareddy *et al.* (2006)), and femurs (Zafar *et al.* (2004)).This impeded ovariectomy-induced loss of bone structure and thus improved bone health (Scholz-Ahrens *et al.* (2002)). It was seen that consumption of a mixture of prebiotic fructo-oligosaccharides by pubertal adolescents elevated calcium absorption, enhanced bone mineralization, (Abrams *et al.* (2005),

Griffin *et al.* (2002)) and led to an increase in whole-body BMC and BMD (Abrams *et al.* (2005)). In another study, consumption of a product rich in transgalactooligosaccharides (TOS) caused an increase in calcium absorption in postmenopausal women (van den Heuvel *et al.* (2000)).

2.4.1 Effect of SCF

Soluble Corn Fiber (or Soluble Maize Fiber) is a corn-derived non-digestible carbohydrate which has been shown to have beneficial effects on bone and human health (Whisner *et al.* (2016)). In previous studies, a Soluble Corn Fiber(SCF) supplemented diet consumed by rats (Knapp *et al.* (2013)) and humans (Klosterbuer *et al.* (2013)), increased SCFA production (Bassaganya-Riera *et al.* (2011)), and showed a positive influence on the microbial community of the gut (Knapp *et al.* (2013), Klosterbuer *et al.* (2013)). A comparison between eight novel prebiotic fibers such as SCF, soluble fiber dextrin, resistant starch (RS), pullulan, and polydextrose, was conducted on weanling rat models (Weaver *et al.* (2010)). It was observed that SCF tended to increase bone mineral density, bone mineral content, cortical thickness, cortical area, and peak breaking strength of femur. SCF also resulted in an increase of total SCFA production and faecal content weight (Weaver *et al.* (2010)). It was demonstrated that consuming SCF(12 g/d) increased calcium absorption efficiency by 12% in a heterogenous population, of 24 adolescent girls and boys (12-15 years), during two three-week sessions of controlled feeding study (Whisner *et al.* (2014)). This was the first study which linked a diet-induced change in the gut microbiota with an increase in calcium intake in healthy individuals (Whisner *et al.* (2016)).

In another related study, 10 or 20 g fiber/d from PROMITOR SCF 85 was added to an uncontrolled diet of a homogenous population including adolescent girls (aged 11-14 years) over a 30-day period. This increased calcium absorption by 13.3% and 12.9% for

10 and 20 g fiber/d from SCF, respectively (Whisner *et al.* (2016)).

Despite the differences in study designs, both the above experiments showed a positive influence on calcium absorption and minimal gastrointestinal symptoms, thereby supporting the effectiveness of this prebiotic fiber in improving bone health (Whisner *et al.* (2016), Whisner *et al.* (2014)).

Further work to understand the exact mechanism by which SCF affects intestinal microbiota and calcium absorption is required. More studies are needed to analyze the long-term effect of dietary fibers on bone density and strength ((Whisner *et al.* (2016)).

2.5 Important Gut Microbes

The human gut microbiome is a stable and diverse environment which is dominated by bacteria from the phyla *Bacteriodetes* and *Firmicutes* (Clarke *et al.* (2014)). The intestinal bacteria and other microorganisms coexist in a dynamic ecological equilibrium (Martin *et al.* (2010)). Most of these bacterial species belong to the genera, *Bacteroides, Clostridium, Lactobacillus, Eubacterium, Faecalibacterium, Bifidobacterium* (Martin *et al.* (2010)), *Ruminococcus, Peptococcus, Peptostreptococcus* (Guarner and Malagelada (2003)). There is also a high abundance of the genera *Prevotella* and *Ruminococcus* which is independent of body mass index, age, or gender (Clarke *et al.* (2014)). Other genera, such as *Escherichia* and *Lactobacillus*, are present to a lesser extent (Guarner and Malagelada (2003)).

The bacterial enterotypes within the gut ecosystem exhibit differential functional capabilities. The genus *Bacteroides* plays a very important role in the functioning of the host as it alone constitutes about 30% of all bacteria in the gut (Sears (2005)). It was seen in a controlled feeding study that the *Bacteroides* and *Prevotella* enterotypes are associated with protein/animal fat and carbohydrate rich diets respectively (Clarke *et al.* (2014), Morgan *et al.* (2013)). Some of the genera of bacteria that are most efficient

in producing SCFAs include *Bacteroides, Bifidobacterium, Eubacterium, Lactobacillus, Clostridium, Roseburia, and Prevotella* (Clarke *et al.* (2014), Neish (2009)).

Two recent studies on adolescents by Whisner *et al.* (2016) and Whisner *et al.* (2014), showed that the changes in the gut microbiota are correlated with improvements in calcium absorption. It was seen that the consumption of SCF was associated with an increase in the proportion of microbes from the phylum *Bacteroidetes* as compared to the treatments without SCF, and decrease in the proportions of bacteria from the phylum *Firmicutes*.

The proportions of bacterial genera *Actinomyces* and *Pseudomonas* of the phylum *Actinobacteria* and other *Erysipelotrichaceae* of the phylum *Firmicutes* decreased as calcium absorption with the SCF treatment increased. Negative correlations were also observed on genera *Paraprevotella, Megamonas, Sutterella, Parabacteroides*, other *Bacteroidales*, and other *Clostridiaceae* (Whisner *et al.* (2016), Whisner *et al.* (2014)).

The proportions of bacterial genera *Parabacteroides, Bifidobacterium*, unclassified *Lachnospiraceae, Dialister, Bacteroides, Butyricicoccus, Oscillibacter*, and *Clostridium* increased after the addition of SCF to the diet as compared to the diet without SCF. This indicates that these microbes are involved in SCF fermentation (Whisner *et al.* (2016), Whisner *et al.* (2014)). The phylum *Bacteroidetes* includes a large number of starch fermenting bacteria. *Bifidobacterium* species are known for their health benefits to the host and are commonly used as probiotics (Whisner *et al.* (2016)).

Therefore, the role of the microbiome in fermentation and calcium absorption and hence its influence on bone health, is a complex mechanism and not mediated by a single species (Whisner *et al.* (2016)).

Chapter 3

BACKGROUND ON MACHINE LEARNING TECHNIQUES

Microbiome research generates large scale data with hundreds of samples and large number of features which allows the use of sophisticated analysis methods such as machine learning algorithms to derive functional relationships between the microbiome and properties of various ecosystems (Namkung (2020)).

3.1 Machine Learning in Microbiome Studies

Machine learning algorithms are computational and statistical data analysis methods, used to build and adapt models on data in order to draw inferences, predict patterns, and learn new tasks. Machine learning methods are categorized into two groups; supervised and unsupervised learning. Supervised learning is used to build a model which can explain the relationship between input and output variables. Unsupervised learning, on the other hand, is used to explore unknown patterns or structure of a given data (Namkung (2020)).

Supervised learning methods are used to identify the relationship between microbiome profiles and host traits. The traits of the dataset are known, and a model is trained to recognize feature characteristics associated with the trait (Zhou and Gallins (2019)). Supervised learning approaches are classified into classification and regression based on the type of trait. Classification predicts the class to which the sample belongs and regression predicts the value of the trait.

A few supervised learning methods that are most commonly used for microbiome host trait predictions, are lasso, ridge, and penalized regression, Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), k- Nearest Neighbors(k-NN), Ran-

dom Forest (RF), and Gradient Boosting. Linear models such as lasso and ridge regression allow simple fitting of continuous variables as a function of feature vectors (Zhou and Gallins (2019)).

Support vector machine (SVM) is a classification method which finds a hyperplane in a high dimensional space, which is a margin between two samples of classes of training data points [ref 36]. Linear Discriminant Analysis (LDA) is a method used to find a linear combination of features that separates two or more classes of objects. k-NN can be used for both classification and regression predictive problems. It is a model that classifies data points based on the points that are most similar to it. Random forest is an ensemble learning method which works on the concept of bagging, where multiple trees are built on bootstrap samples and a random subset of variables, and the outcome is obtained by averaging prediction values or voting for a specific outcome. Gradient boosting is a technique which produces an ensemble model of weak prediction models such as decision trees. The model is built successively by computing weights for the individual trees and optimizing a differentiable loss function (Zhou and Gallins (2019)).

The structure of the human microbiome differs widely among individuals and this makes it difficult to use traditional statistical models to identify populations of microbes that are associated with disease. Traditional statistical approaches also consider the effect of bacterial population individually and do not account for the variation in the human microbiome (Yazdani *et al.* (2016)).

Therefore, machine learning (ML) models are being used in recent studies because they can be used to effectively account for the interpersonal microbiome variation by considering the relative abundance of each bacterial population in the context of other bacterial populations (Topçuoğlu *et al.* (2020)). Machine learning methods are also used to reduce the amount of time required to manually investigate huge amounts of

data obtained from metagenomic sequencing (Yazdani *et al.* (2016)).

In previous studies, machine learning methods were widely used to understand the composition of the microbiome and how it plays a role in the host health. Johnson *et al.* (2016) compared multiple machine learning methods for regression analysis between microbiome profile and post-mortem intervals such as SVM regression, k-NN, lasso/elastic-net regression, and random forest. Random forest, SVM, and Lasso logistic regression methods have also been used to identify microbiome profiles that are associated with specific diseases such as obesity (Le Chatelier *et al.* (2013)), colorectal cancer (Zeller *et al.* (2014)), and Irritable Bowels Syndrome (IBS) (Fukui *et al.* (2020)). Machine learning techniques were also used to discover Type II diabetes associated with gut microbiome profiles (Qin *et al.* (2012)).

In some studies, it was seen that the random forest model outperformed other machine learning methods including linear discriminant analysis (LDA), quantitative discriminant analysis (QDA), k-nearest neighbor (kNN), and support vector machine (SVM) classifiers (Meding *et al.* (2012), Namkung (2020)). In the study by, Topçuoğlu *et al.* (2020), three linear models, L2-regularized logistic regression, L1- and L2- regularized SVMs with linear kernel, and four nonlinear models, SVM with radial basis function kernel, decision tree, random forest, and gradient boosted trees were trained and evaluated using faecal 16S rRNA sequence data to predict the presence of colonic screen relevant neoplasias (SRNs). It was observed that the predictive performance of the random forest model was higher than other ML models. It was demonstrated that the most complex model need not necessarily perform the best and the most interpretable models performed nearly as well as the nonlinear models (Topçuoğlu *et al.* (2020)).

In studies using machine learning models to perform statistical analysis, the data was split into training and test sets (usually in the ratio of 80:20) (Topçuoğlu *et al.*

(2020)), hyperparameters of the model were selected using repeated k-fold CV on the training set, the model with these hyperparameters was trained on the full training set and applied to the held-out data to evaluate the predictive performance of the model. This process was repeated multiple times to obtain a robust robust interpretation of model performance (Namkung (2020), Topçuoğlu *et al.* (2020)). The performance of regression models is evaluated using Pearson's correlation coefficient or Root Mean Squared Error (RMSE), and Mean Absolute Error(MAE). RMSE is used more often as it gives more weight to larger error and is inter- pretable in units of original response variables (Namkung (2020)).

When machine learning methods are applied to data in the microbiome field, certain problems such as, determining which methods have to be used and how they are implemented, evaluating models using the entire dataset without setting aside test data, variation between cross-validation and testing performances, and variation between the predictive performance on different folds of cross-validation, arise. Therefore further work is needed to improve reproducibility and reduce the overestimation of model performance (Topçuoğlu *et al.* (2020)).

3.2 Random Forests for High-Dimensional Data

A High-dimensional dataset is the case where the number of features is larger than the number of observations or samples. Even though it is possible to perfectly fit the training data in the high-dimensional setting, the resulting linear model will perform extremely poorly on an independent test set, and therefore does not constitute a useful model (James *et al.* (2013)). Therefore non-linear models, such as RF, are preferred.

It was seen in previous studies that the RF model demonstrated a higher performance in comparison with other algorithms with respect to high dimension and low sample data sets (Guo *et al.* (2010), Gunduz and Fokoué (2015), Luan *et al.* (2020)).

When building each tree in the random forest model, roughly about one third of the training set is not used. The fraction of training data which is left out certainly contains some outliers, and each outlier has a probability of e^{-1} of not affecting the base learner. Therefore, it can be deduced that by averaging this exclusion of outliers over many replications, a robust model can be achieved through bootstrap aggregation. In the Random Forest model, the extremely high dimensionality of the data is addressed by variable selection performed by random subspace learning. The effect of outliers is reduced or eliminated by subsampling,these features attribute to the overall superior performance of Random Forest (Gunduz and Fokoué (2015)).

In another experiment it was seen that the RF model was robust to limited data, and showed an acceptable predictive performance at a low sample size. The RF model enabled adequate capture of the information from the data and partially reduced the uncertainty of the model predictions for small datasets (Luan *et al.* (2020)).

It also provides a measure of the prediction power of individual variables called variable importance which can be used to select a few key features for further study (Topçuoğlu *et al.* (2020)). Random Forest algorithm is chosen because it is scale invariant, non-linear, and robust to outliers, missing values, and overfitting (Yazdani *et al.* (2016)).

Chapter 4

EXPERIMENT DESIGN

The dataset used for this thesis is based on the study conducted by Whisner *et al.* (2014)

4.1 Subjects

The subjects participating in this study were healthy adolescent girls and boys recruited from local schools, community centres and Indiana extension offices and also by sending direct mails to the surrounding areas. A total of twenty-four adolescents including, fifteen boys, aged 13-15 years, and nine girls, aged 12-14 years, took part in the metabolic studies. The participants were ethnically diverse with a distribution of eleven Asian, six Hispanic, one Black and six multi-racial teenagers.

A 6-day diet record was used to assess their habitual dietary intake and questionnaires based on brief medical history, maturational age, and physical activity were used to screen the participants to determine their eligibility.

The criteria for exclusion included abnormal liver or kidney function, malabsorptive disorders, anaemia, history of using medications influencing Ca metabolism, body weight outside the 5^{th}-95^{th} Body Mass Index(BMI) percentile for age, regular consumption of illegal drugs, non-prescription drugs, or any kind of contraceptives, and pregnancy.

The subjects were not permitted to take any nutritional supplements while participating in these studies. The study was conducted according to the guidelines laid down in the Declaration of Helsinki, and all procedures involving human subjects were approved by the Institutional Review Board of Purdue University. A written informed consent was acquired from all the subjects participating in this study (Whisner *et al.*

(2014)).

During the first session of the camp, anthropometric parameters including weight, sitting height, waist circumference and hip circumference were measured. Standing height was measured at the beginning of the first session using a wall-mounted stadiometer, and to ensure that the weight remained stable throughout the sessions, it was monitored each morning with an electronic digital scale.

Among all the participants, three members did not undergo the fractional Ca absorption test in both sessions and one member attended only one of the camp sessions. Therefore, analyses included twenty-three participants in all (Whisner *et al.* (2014)). The characteristics of the participants are shown in Table 4.1 (Whisner *et al.* (2014)).

Characteristics	Females		Males	
	Mean	SD	Mean	SD
Age(years)	13.3	1.0	13.5	0.9
Weight(kg)	59.9	13.2	61.1	11.8
BMI(kg/m^2)	24.1	4.0	22.4	3.1
Total Body BMD(g/cm^2)	1.07	0.11	1.04	0.11
Total Body BMC(g)	2115	329	2316	424
Total Spine BMD(g/cm^2)	1.09	0.13	1.04	0.14
Femoral Neck BMD(g/cm^2)	1.03	0.18	1.05	0.15

Table 4.1: Characteristics of Participants. BMI: Body Mass Index, BMD: Bone Mineral Density, SD: Standard Deviation

4.2 Study Design

The studies were designed to mimic the experience of a summer camp, where adolescent boys and girls were taken on field trips, and were allowed to participate in a

variety of recreational and educational activities. During the duration of the camp, the participants were housed in University Residence Halls at Purdue University.

For this experiment, a double-blind, cross-over design was used in which the participants received two treatments in a randomised order. One treatment was labelled SCF treatment where a 12 g/d SCF dose was provided and the other treatment was labelled Control Treatment (CON) where a 0 g/d SCF dose was administered. The study involved two 3-week balance studies which were separated by a 7-day washout period (Whisner *et al.* (2014)).

4.3 Diets

The diet provided, during both the 3-week sessions, consisted of foods typically eaten by adolescent children such as spaghetti, hamburgers, sandwiches and potato chips and was a controlled diet.

The diets were designed in such a way that body weight was maintained and constant amounts of key nutrients were present. The controlled diets were provided as a 4 day cycle menu with three meals and two snacks daily. SCF was present in Welch's® fruit snacks (WELCH Foods, Inc.) and was provided at lunch and dinner, divided into two 0 or 6 g fibre doses. The diets contained 14% protein, 33% fat, 53% carbohydrate, 5mg vitamin D, 1100mg P, 2300mg Na and 600mg Ca, on an average.

The SCF ingredient called, PROMITOR® Soluble Corn Fiber 70, which was provided by Tate & Lyle is a fermentable, non-digestible carbohydrate containing a minimum of 70% soluble dietary fibre.The basal diet contained 15g of fibre and the intervention product contributed an additional 0g or 12g SCF. This yielded a dietary fibre content of 15g and 27g, in total, for the CON and SCF treatments, respectively (Whisner *et al.* (2014)).

4.4 Bone Health Measurements

Bone mineral content and bone mineral density were measured using a method called Dualenergy X-ray absorptiometry (GE Lunar) in order to determine the bone status of the participants. Dual-energy X-ray Absorptiometry (DXA) is one of the standard methods to measure Bone Mineral Density (BMD) using spectral imaging. The images are processed to compute the Bone Mineral Content (BMC) per projected area which is referred to as the projected BMD. BMC is calculated by summing the BMD values over the projected area.

BMD is the amount of bone mineral present in the bone tissue. Bone mineral content (BMC) is a measurement of bone mineral found in a specific area and is measured in grams (g). BMD is measured in grams per centimeter squared (g/cm^2). BMD is calculated by dividing BMC by area. During one of the sessions, DXA scans were performed to collect bone measurements of the whole body, spine, forearm and both hips. For this study, the following measures of bone health were collected from each participant, total body BMD, hip BMD, spine BMD, total body BMC, hip BMC, and spine BMC (Whisner *et al.* (2014)).

4.5 Faecal Sample Processing and Microbial Community Composition

The composition of the faecal microbial community was determined from the samples collected at the beginning and end of each session for every participant. Sterilised double distilled water was added to the frozen faecal samples and the slurries were stored at -20 degree celsius until DNA extraction. DNA was extracted from 50 - 100mg of faecal material using the FastDNA® SPIN Kit for Soil as mentioned in Ariefdjohan *et al.* (2010).

Trained counsellors supervised the participants during activity, meal and sample

collection periods for 24 h each day. Unconsumed food from meals was collected and its amount recorded (Whisner *et al.* (2014)). The presence of stomach noises, flatulence, bloating and abdominal pain among the participants was evaluated daily using a short questionnaire.

The phylogenetic diversity of bacterial communities was determined using *16S* ribosomal RNA (rRNA) gene sequences obtained using 454 FLX titanium chemistry and Roche Genome Sequencer and primers that amplify the V3–V5 region of the *16S* rRNA gene as described in (Nossa *et al.* (2010)). The *16S* rRNA gene sequence dataset was analysed using the Quantitative Insights Into Microbial Ecology (QIIME) pipeline as described in Caporaso *et al.* (2010). The pipeline uses a multi-software approach to perform quality filtering, operational taxonomic unit (OTU) picking, taxonomic assignment, alpha diversity (bacterial diversity within each sample) and beta diversity (bacterial similarities and differences among the samples) measures. (Whisner *et al.* (2014), Nossa *et al.* (2010)]. The representative OTU was given final taxonomic assignments using the Greengenes dataset and Ribosomal Database Project (RDP) classifier at a confidence of 80% (Whisner *et al.* (2014)).

Taxonomy is known as the science of defining and naming groups of biological organisms based on certain shared characteristics. The evolutionary relationship among the microbes represented by each operational taxonomic unit, is defined by the taxonomy. In this study the taxonomy order used, from general to specific, is as follows: kingdom *(k)*, phylum *(p)*, class *(c)*, order *(o)*, family *(f)*, and genus *(g)*. *Firmicutes* was the most dominant phylum followed by the phyla *Bacteroidetes*, *Actinobacteria*, and *Proteobacteria* (Whisner *et al.* (2014)).

The microbial community varies with the host's health condition, age, and diet. The composition is different in the stomach compared to that of the colon owing to the space and nutrients available in the large intestine. A major source of intestinal

metabolism comes from the processing of dietary nutrients by gut microbes. Therefore, by assessing the metabolic changes in feces, nutrient microbiota relationships can be studied. In clinical and preclinical trials, fecal metabolic monitoring should be considered to explore how host metabolism is impacted by dietary habits through metabolic activity of bacteria (Martin *et al.* (2010)).

In this study a total of 178 microbes were identified and the GM sequence constitutes values which represent the proportions of all the microbe genera present in the faecal samples collected from the participants (Whisner *et al.* (2014)).

Chapter 5

STATISTICAL ANALYSIS

In this project, the main aim was to establish a relationship between the gut microbiome and bone health. The analysis of this relationship was carried out in 2 steps. First, only the sequences from the 0 g/d SCF dose were used. Second, the sequences from both the 0 g/d SCF and 12 g/d SCF doses were used to understand the effect of SCF.

5.1 Problem Formulation

During the two phases (0 g/d SCF and 12 g/d SCF) of this study, faecal samples were collected from each of the twenty-three participants and used for sequencing to determine microbe sequences. A total of 178 microbe sequences were recorded for each participant. DXA method was used to record six different measures of bone health during the two phases, including the following, Total Body Bone Mineral Density (TBBMD), Spine Bone Mineral Density (SPBMD), Hip Bone Mineral Density (HPBMD), Total Body Bone Mineral Content (TBBMC), Spine Bone Mineral Content (SPBMC), and Hip Bone Mineral Content (HPBMC). This constitutes a high-dimensional low-sample dataset, as the number of features exceeds the number of samples.

To understand the relationship between the microbes in the gut and the measures of bone health, the GM sequence, consisting of the 178 microbe proportions, was used to predict the measures of BMD and BMC in the two phases of the study, using the regression machine learning algorithm.

In statistical modeling, regression analysis is a collection of statistical processes which are used to estimate relationships between a dependent variable (outcome vari-

able) and one or more independent variables, also known as predictors or features. Regression analysis is widely used for the purpose of prediction and forecasting. It can also be used to infer causal relationships between the independent and dependent variables.

Regression models consist of the following components:

1. Unknown parameters, often denoted as a scalar or vector β.
2. Independent variables, denoted as a vector X_i (where i denotes a row of data).
3. Dependent variable, denoted using the scalar Y_i
4. Error terms, denoted using the scalar e_i.

It is proposed that, Y_i is a function of X_i and β, with e_i which represents an additive error term,

$$Y_i = f(X_i, \beta) + e_i \tag{5.1}$$

In this algorithm, the independent variables are represented by the feature vector set formed by 178 GM sequences collected from 23 participants, from two phases. The dependent variable was represented by one of the measures of bone health mentioned above.

The goal is to estimate the function $f(X_i, \beta)$ that comes very close to fitting the data. The regression method ultimately provides an estimate of β , denoted by $\hat{\beta}$ to distinguish the estimate from the true parameter value that generated the data. The fitted value,

$$\hat{Y}_i = f(X_i, \hat{\beta}) \tag{5.2}$$

can then be used for prediction purposes or to assess the accuracy of the model in explaining the data.

5.2 Prediction Analysis

As mentioned earlier, many recent studies have used machine learning regression and classification algorithms to analyze the gut microbiome data. Ensemble learning methods, especially Random Forest, have been widely used and are effective in the case of high dimensional data.

In this project, the Random Forest regression algorithm was used as the main model to predict the measures of BMD and BMC. Another ensemble regression model, Extremely Randomized Trees, was also built in order to validate the predictions. The ensemble learning algorithms that were used are explained below.

5.2.1 Random Forest

Ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the individual learning algorithms considered separately.

Random forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputs the class that is the mode of the classes, in case of classification or the mean prediction of the individual trees, in regression.

Decision Tree

Decision trees are models which are built by learning a hierarchy of if/else questions on the given data, leading to a final decision. A decision tree is constructed by splitting the source dataset into two or more homogeneous subsets based on a set of splitting rules focused on the input variables. The technique of recursive partitioning is used where each derived subset is further split in a recursive manner. This splitting process is terminated when the subset obtained at a node contains similar values of the target variable, or when the splitting ceases to add value to the predictions made. This

process is called top-down induction of decision trees (TDIDT) and it is an example of a greedy algorithm. The structure of the decision tree is shown in Figure 5.1.

The Root Node represents the entire sample and further gets divided into subsets. The Interior/Decision Nodes represent the features of a data set and the branches represent the decision rules. The Leaf/Terminal Nodes represent the final outcome and do not split further. The final prediction is given by the average of all the values of the dependent variable in that specific leaf node. The tree predicts the final value by performing multiple iterations.

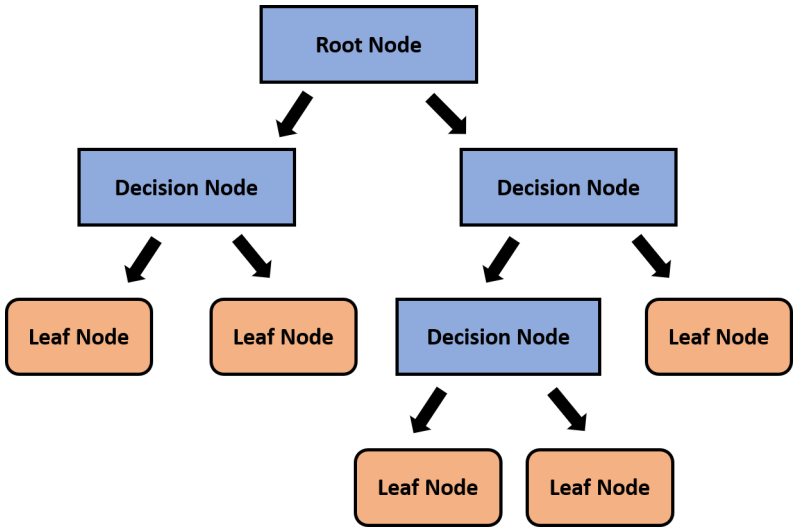


Figure 5.1: Decision Tree

One of the main drawbacks of decision trees is that they tend to overfit the training data, and lead to a low bias and high variance condition.

Bootstrap aggregation, or bagging, is a procedure used for reducing the variance of a statistical learning method (James *et al.* (2013)). The Random Forest training algorithm applies the technique of bootstrap aggregating, or bagging, to the individual trees.

Bagging

Given a training set, $X = x_1, ..., x_n$, and the corresponding response variables, $Y =$

$y_1, \ldots y_n$, the bagging algorithm selects a random subset of this training set by replacing the samples in each iteration, and fits the trees to this subset. The algorithm works as follows, For each iteration $b = 1,...,B$:

- n training examples from X, Y are sampled with replacement, called X_b, Y_b.
- A classification or regression tree f_b is trained on X_b, Y_b.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x') \tag{5.3}$$

This bootstrapping procedure decreases the variance of the model, without increasing the bias and this leads to a better model performance. This implies that, as long as the trees are not correlated with one another, the average prediction of many trees is not sensitive to noise even though the predictions of a single tree can be highly sensitive to noise.

Random Forest

In the Random Forest algorithm, multiple deep decision trees are trained on different parts of the same training set (bootstrapping), and the predictions of all the trees are averaged as shown in Figure 5.2. Each individual tree has high variance, but low bias. Averaging these trees reduces the overall variance. This is achieved at the expense of a small increase in the bias and reduces interpretability, but boosts the overall performance.

Random forests improve the variance reduction of bagging by reducing the correlation between the trees, without increasing the variance too much (Hastie *et al.* (2009)). During the process of constructing decision trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set

of p predictors, in this case, the 178 GM sequences. Only one of those m predictors are used to make a decision at the split. In general, the number of predictors considered at each split is approximately equal to the square root of the total number of predictors (James *et al.* (2013)). The number of decision trees, the depth of each decision tree, the number of samples and predictors to be considered at each split, are some of the parameters that can be adjusted in order to obtain a good model.

The main difference between Bagging and Random Forests is the choice of predictor subset size m. If multiple decision trees are built using all the available predictors, i.e., $m = p$, then it is called bagging. (James *et al.* (2013)).

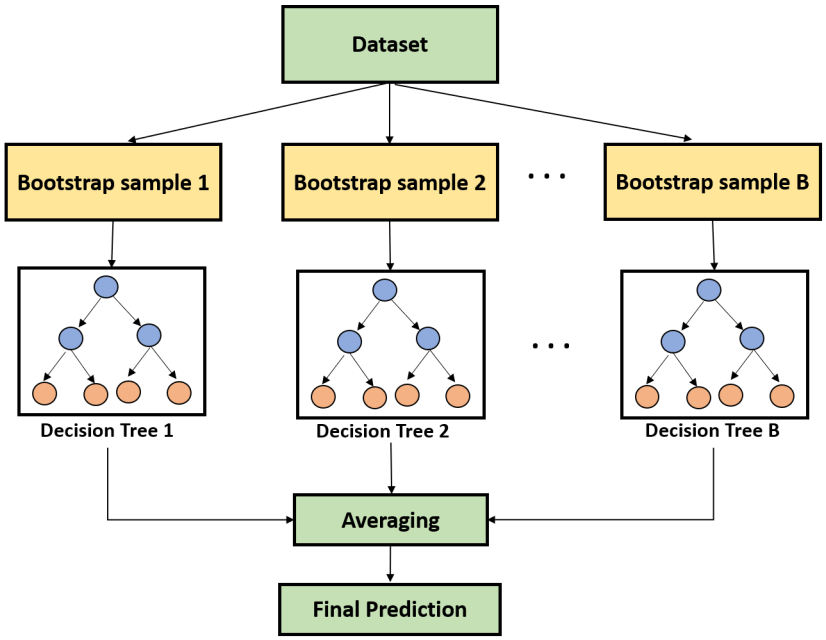


Figure 5.2: Random Forest Algorithm

The Random Forest algorithm (Hastie *et al.* (2009)) can be summarized as follows:

- For $b = 1$ to B:

  (a) A random bootstrap sample of size N is drawn from the training data .

  (b) A random-forest tree T_b is built on the bootstrapped data, by repeating the

following steps in each iteration, at the terminal node of the tree, until the minimum node size is reached,

i. Selecting m variables from the p variables at random

ii. Among these m variables, picking the best variable/split-point

iii. Splitting the node into two daughter nodes

- Generate an ensemble of trees represented by $\{T_b\}_1^B$.
- The ensemble of trees can be used to make a prediction on a new unseen point x by averaging the predictions of all the trees in the forest given by,

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{5.4}$$

5.2.2 Extremely Randomized Trees

Extremely randomized trees are an ensemble of individual trees, which are very similar to random forests. As in the RF model, the number of features to be considered at each node are selected at random. However, in extremely randomized trees, the splits are computed in a random fashion and each tree is trained using the entire learning sample instead of using a bootstrap sample.

For each feature under consideration, a random cut-point is selected to divide the parent node into two child nodes, instead of calculating the best split at that node. This threshold value is selected from a uniform distribution within the feature's empirical range. The split that yields the highest score amongst all the randomly generated splits is chosen as the final split at that node. In comparison to the RF algorithm, this method reduces the variance of the model to a greater extent but at the expense of a slightly greater increase in bias.

In the next sections, the process to develop the regression model to predict the bone health measures is highlighted.

5.3 Model Selection

Model selection is the process of selecting a machine learning model, from among a collection of different machine learning models and also across models of the same type configured using different model hyperparameters, for a given training dataset. The performance of different models is measured in order to choose the best model. Once the final model is chosen, it is assessed by estimating its prediction error on the test data (Hastie *et al.* (2009)).

The analysis of the relationship between gut microbiome and bone health was carried out in 2 steps. First, only the sequences from the 0 g/d SCF dose were used. The independent variable set was represented by 23 samples and 178 features. Second, the sequences from both the 0 g/d SCF and 12 g/d SCF doses were used to understand the effect of SCF. The independent variable set included 46 samples in all, and 178 features. In both the cases the dependent variable to be predicted was one of the six bone strength measures which included 23 samples and 46 samples, respectively.

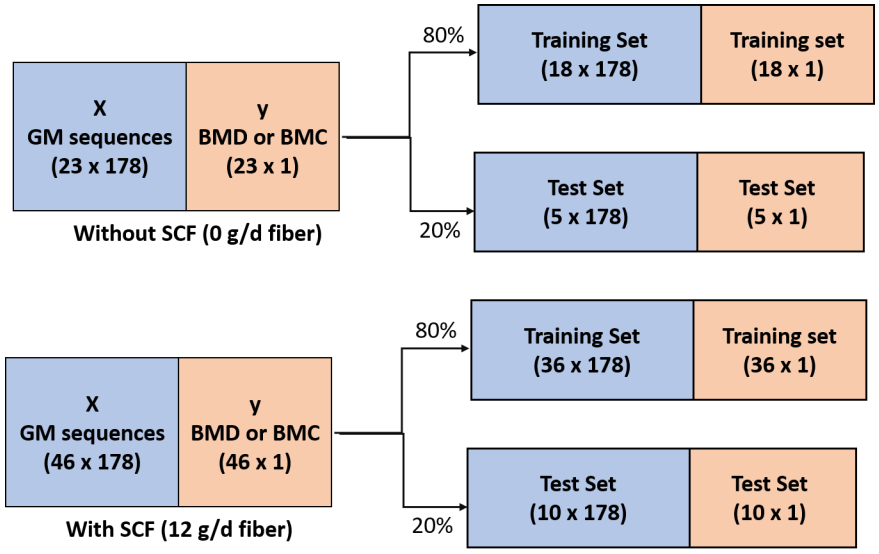


Figure 5.3: Dataset Representation

The training dataset is a set of examples which is used during the learning process

and is used to fit the parameters of the model. The validation dataset or development set is a set of examples used to tune the hyperparameters of the model and it provides an unbiased evaluation of a model fit on the training dataset. Different models are trained on the training data set by minimizing an appropriate error function. The performance of the models is then compared by evaluating the error function using an independent validation set, and the model which has the smallest error with respect to the validation set is selected. Finally, the test dataset, which is independent of the training set, is a dataset used to provide an unbiased evaluation of a final model fit on the training dataset. Minimal overfitting takes place if the final model selected, fits the test dataset well.

However, when the data is partitioned into three subsections, the number of samples which can be used by the model for learning drastically reduces. One way to overcome this problem is to use a procedure called cross-validation. It is required that a test set should still be held out for final evaluation, but the validation set is no longer necessary. Cross-validation is especially useful when the training dataset is very small and holding out part of the data just for validation purposes is not affordable (Zheng (2015)).

In the first case (0 g/d SCF dose), I split the above mentioned dataset into a training set and test set in the ratio of 80-20. The training set included 80 percent of the samples (18 sequences) and the test set included the remaining 20 percent of the samples (5 sequences). In the second case (12 g/d dose), the dataset was split into a training set of 36 samples and test set of 10 samples. This is depicted in Figure 5.3.

I used a Random Forest (RF) regression model to train the training set samples in each case. In order to optimize the model, I tuned the hyperparameters of the RF model using the procedure described below.

5.3.1 Hyperparameter Tuning

Hyperparameter optimization is the process of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process. A tuple of hyperparameters is found that yields an optimal model which minimizes a predefined loss function on given independent data.

The hyperparameter values are changed when running a learning algorithm over the training set and this results in a set of different models. Model selection refers to the process of finding the model which gives best performance from this set of models (Raschka (2018)). The process of hyperparameter tuning and model selection are performed on the training set simultaneously (Raschka (2018)). The learning algorithm optimizes an objective function on the training set along with hyperparameter tuning. After the tuning stage, a reasonable approach is to select a model based on the test set performance. However, using the test set again and again would introduce a bias and result in overly optimistic estimates of the generalization performance leading to the leakage of information from the test set. To avoid this problem, a three-way split can be performed by dividing the dataset into training, validation, and test datasets (Raschka (2018)). In this way the training set can be used to fit the models, the validation set can be used to estimate the prediction error for the selected model, and this pair can be used for hyperparameter tuning . The test set is used to assess the generalization error of the final model (Hastie *et al.* (2009)).

K-fold cross-validation is a special case of cross-validation where the training set is split into k smaller sets. For each of the k folds generated, k-1 of the folds are used as training data to train a model and this resulting model is validated on the remaining part of the data.

The procedure to tune hyperparameters for model selection outlined in Figure 5.4, is described as follows:

- Step 1: The dataset is split into two parts, a training and an independent test set. The test set is kept aside for the final model evaluation step. In this case, I randomly selected 20 percent of the data as the test set.

- Step 2: Various hyperparameter settings can be experimented with, by using techniques such as Bayesian optimization, Randomized search, or Grid search. The k-fold cross-validation method is applied on the training set, resulting in multiple models and performance estimates (Raschka (2018)). The overall performance is considered to be the mean of the performance on all k folds (Zheng (2015)).

- Step 3: The hyperparameter combination that gave the best results in the k-fold cross validation procedure was selected and a final RF model was fit with these values, on the entire training set of 23 samples or 46 samples in cases 1 and 2 respectively.

- Step 4: The 20 percent independent test set which was withheld earlier is now used to evaluate the final model obtained from Step 3. The RF model was fit on the test set to obtain the predictions of the bone health measures.

- Step 5: After evaluation, a model can be fit to the entire dataset and this could be the model for deployment. This step is optional.

5.3.2 Grid Search and Random Search

The random forest algorithm has several hyperparameters that can be modified, such as the number of trees, number of observations drawn randomly for each tree,

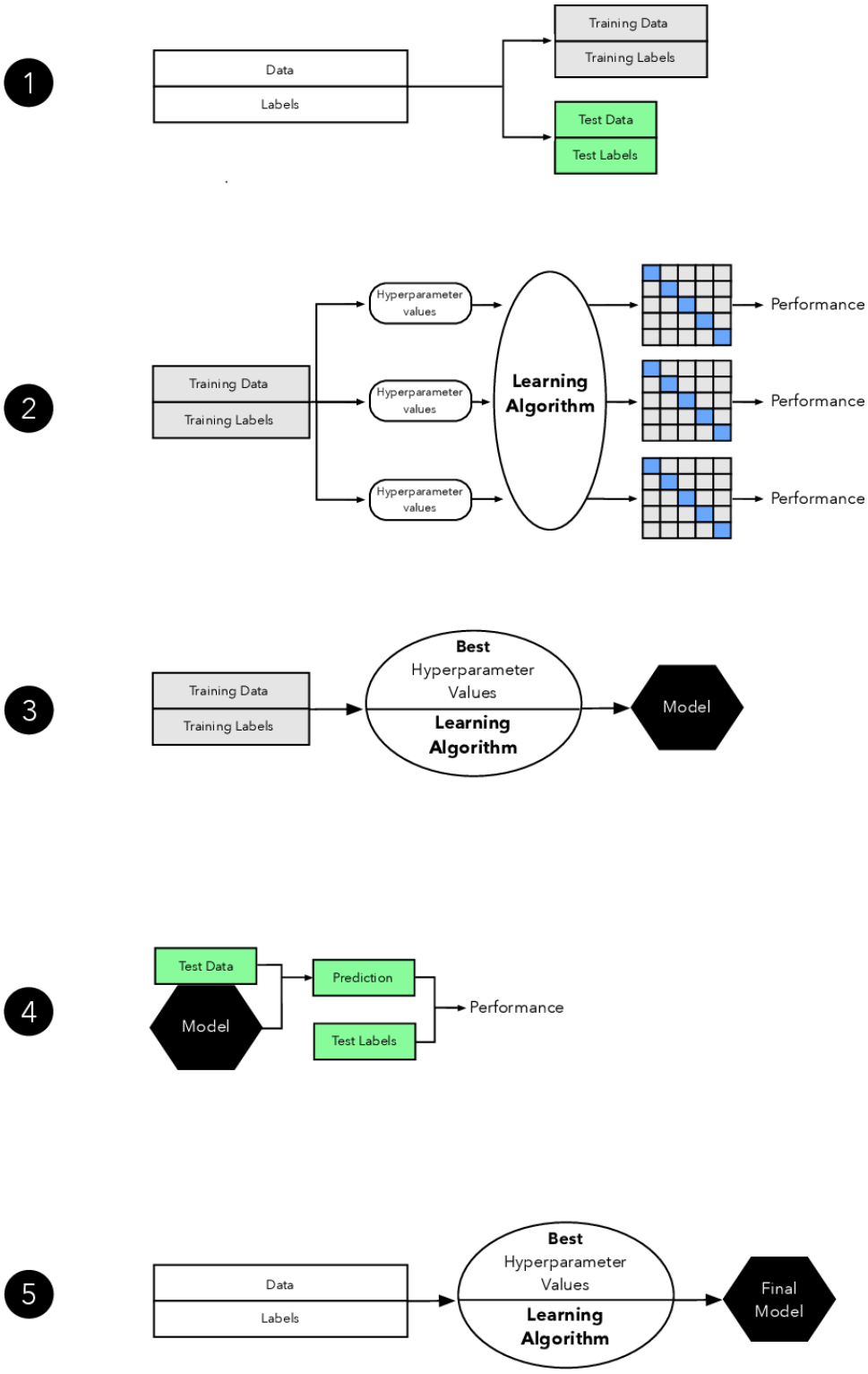


Figure 5.4: Hyperparameter Tuning for Model Selection. From Raschka (2018)

the number of variables drawn randomly for each split, and the minimum number of samples that a node must contain (Probst *et al.* (2019)). One of the major attributes to be tuned is the number of features/variables used in an individual tree. Selecting a small number will reduce the variance of the ensemble but it might also increase the bias of an individual tree in the ensemble. If the dataset comprises a lot of noisy variables, a lesser number of variables selected will decrease the probability of choosing an important variable at a split. Also a higher number of trees gives better performance.

Random Search

Random search uses a randomized search over all the parameters and samples each setting randomly from a distribution over possible parameter values. This technique can be applied to a discrete set of values or it can be generalized to continuous and mixed spaces.

Distributions of each parameter were searched using the Random search technique using 5-fold CV, over 700 iterations. This process narrowed down the range of each parameter enabling trial using specific combination settings in that range, using Grid Search.

Grid Search

Grid-search is used to find the optimal hyperparameters of a model, from a combination of parameters, which results in the most accurate predictions. This is a traditional method of performing hyperparameter optimization and is also known as parameter sweep. In this technique a hyperparameter space of the learning algorithm is specified manually and exhaustive searching of this subset is performed. The grid search algorithm is guided by optimizing a performance metric. Different performance metrics can be measured by evaluation on a held-out test set or cross-validation on the training set.

Grid Search along with 3-fold CV was used in Case 1 (0 g/d fiber) and 5-fold CV in

case 2 (12 g/d fiber), to identify the best hyperparameter settings from a grid of values. The parameters that were tuned include the number of trees in the forest, the number of features assigned to each tree, the depth of each tree, and the number of samples in the node of each tree.

Each hyperparameter setting was run over 100 iterations. In each iteration, one of the k folds is held out as a validation set. A model is trained on the rest of the k – 1 folds and its performance is measured on the held-out fold. This is repeated for all the hyperparameter settings that need to be tested. The best hyperparameter combination was identified as the one which produced the highest performance metric.

5.4 Model Evaluation

The predictive performance of a model is evaluated for the following reasons:

1. To estimate the generalization performance, which is the predictive performance of the model on new data.

2. To increase the predictive performance by tweaking the learning algorithm and selecting the best performing model from a given hypothesis space.

3. To identify the machine learning algorithm that is best-suited for the given problem (Raschka (2018)).

5.4.1 Cross-Validation

Cross-validation is a model validation technique which is used to assess the effectiveness of the results of a statistical analysis on independent data. The aim of cross-validation is to evaluate the model's ability to predict new data, and provide an insight on how the model will generalize to an independent dataset.

A single round of cross-validation involves splitting a sample of data into two sub-

sets, called training set, on which the analysis is performed, and test set, on which validation is performed. In order to reduce variability, multiple rounds of cross-validation can be performed using different partitions, and all the validation results are combined over the rounds to provide an estimate of the model's predictive performance.

Once the best model that fit the training data well was found, its performance was evaluated by fitting it to the 20 percent test set which was set aside initially. In order to see how well the model performs on new data, performance metrics such as Root Mean Squared Error, Mean Absolute Error, Pearson's correlation coefficient, and Spearman's correlation coefficient, between the true and predicted values of the BMD and BMC measures, were calculated.

5.4.2 Performance Metrics

Performance metrics can be used to evaluate how well the model, fit on the training data, performs on unseen test data.

1. Mean Absolute Error (MAE): Absolute error is the difference between the true value and the predicted value. MAE is the average of all absolute errors. It can be is calculated as follows:

$$\frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{5.5}$$

   where y_i is the prediction and x_i is the original value.

2. Root Mean Suared Error (RMSE): It measures the average magnitude of the error. It is calculated as the square root of the average of squared differences between the true and predicted observations given by,

$$\sqrt{\frac{\sum_{i=1}^{n} (y_i - x_i)^2}{n}} \tag{5.6}$$

3. Pearson Correlation Coefficient (r): It is a statistic that measures linear correlation between two variables X and Y. It has a value between +1 and -1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation.

4. Spearman's rank correlation coefficient (ρ): It is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function. The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables. While Pearson's correlation assesses linear relationships, Spearman's correlation assesses whether the relationships are linear or not. A Spearman correlation of +1 or -1 occurs when each of the variables is a perfect monotone function of the other.

In this study, the dataset is very small containing only 23 or 46 samples. In this situation, the generalization performance of the model can be evaluated by using k-fold cross -validation instead of just the test set (Raschka (2018)).

5.4.3 K-Fold Cross-Validation

In k-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data.

The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the validation data. The k results can then be averaged to produce a single estimation. The advantage of this method over repeated random subsampling is that all observations are used for both training and validation, and each

observation is used for validation exactly once.

Typically, given these considerations, one performs k-fold cross-validation using k = 5 or k = 10, as these values have been shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance (James *et al.* (2013)).

The final RF model was evaluated on the entire dataset by splitting the dataset into k-folds, where k=3 in case 1 and k=5 in case 2, as depicted in Figure 5.5. The performance metrics were calculated in each iteration and averaged to obtain the overall value.

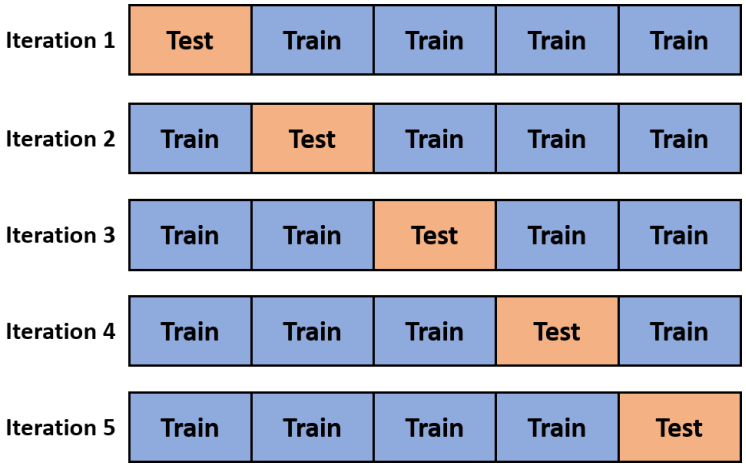


Figure 5.5: K-fold Cross-Validation

The following are the different variations in cross-validation:

- When k = n (the number of observations), k-fold cross-validation is equivalent to leave-one-out cross-validation.
- Repeated k-fold cross-validation: the data is randomly split into k partitions several times. The performance of the model can thereby be averaged over several runs, but this is rarely desirable in practice. The data was split into k-folds in 100 iterations and the performance metrics, described earlier, were calculated by averaging the values over all the iterations.

- Random permutations cross-validation or Shuffle and Split CV: In this method, samples are shuffled and then split into a pair of train and test sets. The number of train/test splits can be specified by the user. This is a good alternative to k-fold cross validation, as it allows a finer control on the number of iterations and the proportion of samples on each side of the train / test split.

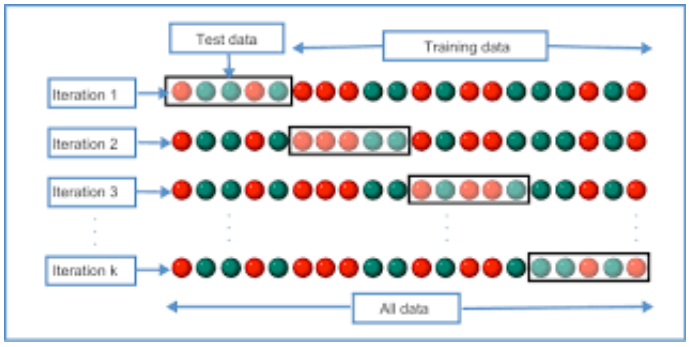


Figure 5.6: Shuffle and Split Cross-Validation. From Wikipedia

To get a better understanding of how the model performs, the entire dataset was shuffled and split into 80-20 percent folds of train/test datasets, in 100 iterations, as shown in Figure 5.6. The RF model was fit on the training fold and evaluated on the test fold by calculating the performance metrics in all iterations. The average of all the iterations was computed to get the final correlation coefficient and other performance metrics.

This process of model selection and evaluation, using the Random Forest algorithm, was used to predict the values of the six bone mineral density and bone mineral content measures. The process was carried out in 2 steps, first using only the samples from the CON treatment (0 g/d fiber) and second using all the samples from the SCF treatment (12 g/d fiber). All the performance metrics obtained in both the phases were tabulated.

The entire procedure was then repeated using the Extremely Randomized Trees regression model, in order to validate the results that were obtained previously using the

Random Forest model.

In order to identify the microbes that played a role in predicting these bone measurements, the permutation feature importance technique was used on the final RF model, to rank the microbes based on their importance.

5.5 Feature Selection

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of features which are most relevant for the construction of machine learning models. Feature selection techniques are used for several reasons:

- Easier interpretation of models by simplifying them
- Shorter training times
- To avoid the curse of dimensionality
- To reduce overfitting and enhance generalization
- To obtain a better understanding of individual features and their relationship to the response variables.

The first part of this project was to predict the measures of bone health (response variable) using the GM sequence consisting of microbe proportions (predictor variables). The aim of this project was to not only make the most accurate predictions of the response variable but to also identify which predictor variables/features (microbes) play the most important in making these predictions. I used the Random Forest model to identify the most relevant microbes which affected the gut-bone relationship.

5.5.1 Feature Importance of Random Forest

The variable importance measures provided by the Random Forest model serve as an effective solution for the problem of feature selection because of the properties of Random Forests such as their ability to model complex interactions, flexibility to work with numerical and categorical variables, good prediction performance, and robustness to noisy variables (Louppe (2014)).

The feature importance values are usually calculated by the mean decrease in impurity mechanism . When decision trees are built in the RF algorithm, the importance of a feature is computed by measuring the effectiveness of the feature at reducing the uncertainty or variance. The value is equal to zero if and only if the variable is irrelevant. However, the drawback of this method is that it shows a bias towards variables with more categories and prefers those variables over others.

5.5.2 Permutation Feature Importance

In this method the importance of a feature is measured by randomly shuffling the values of each feature, without disturbing the other predictor variables or the target variable, and observing how this permutation affects the performance metric of the machine learning model and influences the model performance.

The approach for calculating the permutation feature importance using Random Forest is described as follows:

1. A baseline model is trained on the training set and its evaluation is performed by passing the validation set and obtaining the performance metric. The score recorded in this case is denoted by P_b.

2. The values of one feature are shuffled randomly within that column and this modified dataset is passed to the model to obtain new predictions and a new

performance metric is obtained. The score recorded after shuffling is denoted by P_s. The feature importance is the difference between these two scores.

3. The above procedure is repeated for all features in the dataset.

If performance metrics such as precision, accuracy, or coefficient of determination are used, where larger values are better, the importance is defined as $(P_b - P_s)$. However if error/loss metrics such as log loss or root mean squared error are used, then the importance is defined as $-(P_b - P_s)$. The importance of the feature is measured as the reduction in performance or increase in the model's prediction error after the shuffling has been performed. For example, if shuffling the values of a feature decreases the accuracy, then it is considered as important and it implies that the model relies on that feature for prediction. If the model accuracy remains unchanged, then that feature is not considered by the model for prediction. Therefore no matter which metric is chosen, a higher value implies the feature is more important.The features which show a maximum decrease in accuracy are considered as the most important features.

The results obtained by this permutation mechanism are more reliable as compared to the mean decrease in impurity mechanism, even though it is computationally expensive. In the permutation importance strategy, after permuting each column the model does not have to be retrained. The perturbed test samples can be re-run through the already-trained model, which saves the computation time.

Permutation Feature Importance technique was used along with Shuffle and Split CV as shown in Figure 5.7. The entire dataset was shuffled and split into training and test set folds in the ratio of 80:20. The final RF model was fit on the training fold and the the performance metric (R^2) was calculated on the test fold. Next, the dataset is modified by shuffling each of the feature vector columns. This modified dataset is then passed to the model again to obtain new predictions and new R^2 values. The features

are shuffled 15 times and the average R^2 value is recorded. This process is repeated 100 times and for all the 178 feature vector columns. The final importance value is obtained by taking the average of all the values from all the iterations.

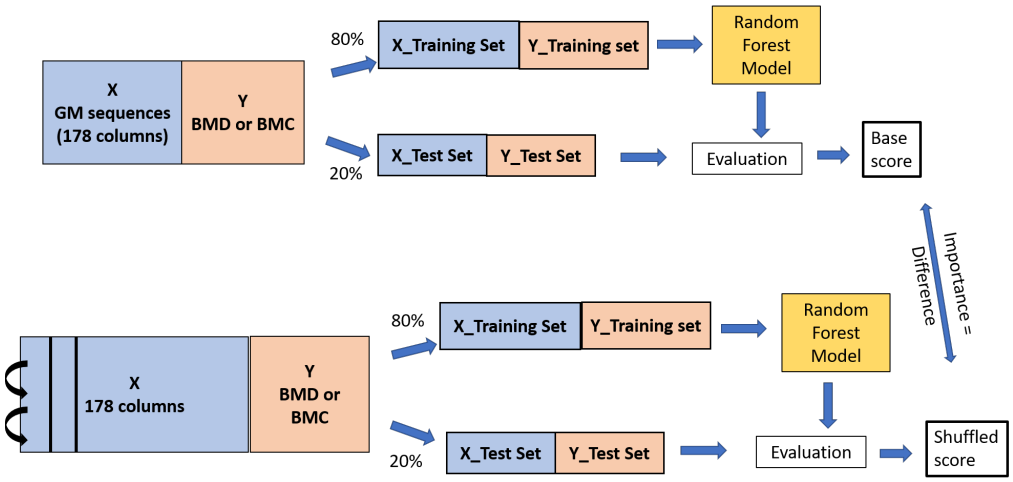


Figure 5.7: Permutation Feature Importance of Random Forest

The entire process of training and evaluating the regression models along with their evaluation and feature selection was implemented using the Spyder (Python 3.7) platform and Jupyter Notebook. The results obtained from model selection, evaluation and feature selection techniques are explained in the next section.

Chapter 6

RESULTS

6.1 Relationship Between Gut Microbiome and Bone Health

The Root Mean Squared Error (RMSE), Pearson's correlation coefficient (r), and Spearman's correlation coefficient (ρ), metrics were calculated between the ground truth values of the bone health measures and the values predicted by the RF model.

The values for both the cases, without and with SCF intervention, can be seen in Table 6.1 and Table 6.2, respectively.

Variable	Range	RMSE	r	ρ
TBBMD	0.86 - 1.24	0.11	0.075	-0.0008
TBBMC	1648.26- 3243.07	434.35	-0.207	-0.17
TSBMD	0.82 - 1.32	0.126	0.35	0.27
TSBMC	31.92 - 78.07	13.62	-0.12	-0.06
HPBMD	0.71 - 1.27	0.165	-0.19	-0.18
HPBMC	3.34 - 6.76	1	-0.4	-0.37

Table 6.1: Performance Metrics for Phase Without SCF (0 g/d Fiber) Using Random Forest Regression Model and Cross-Validation

It can be seen from Tables 6.1 and 6.2 that the correlation coefficient values show an improvement from the CON phase to the SCF phase, in all the six bone measure cases. The Pearson Correlation Coefficient values range between 0.3 and 0.65 and the Spearman Correlation coefficent values range between 0.25 and 0.6, in the case with SCF treatment.

Variable	Range	RMSE	r	ρ
TBBMD	0.86 - 1.24	0.096	0.55	0.49
TBBMC	1648.26- 3243.07	364.96	0.5	0.48
TSBMD	0.82 - 1.32	0.111	0.64	0.56
TSBMC	31.92 - 78.07	11.575	0.55	0.53
HPBMD	0.71 - 1.27	0.137	0.51	0.47
HPBMC	3.34 - 6.76	0.89	0.31	0.26

Table 6.2: Performance Metrics for Phase With SCF (12 g/d Fiber) Using Random Forest Regression Model and Cross-Validation

The performance metrics were also calculated for the values predicted by the Extremely Randomized Trees regression model for the bone health measures, TSBMD and TSBMC, during the two phases, as shown in Table 6.3.

Phase	Variable	RMSE	r	ρ
Without SCF	TSBMD	0.126	0.35	0.29
	TSBMC	13.3	0.044	0.065
With SCF	TSBMD	0.112	0.64	0.56
	TSBMC	10.83	0.59	0.55

Table 6.3: Performance Metrics for Both Phases Using Extremely Randomized Trees Regression and Cross-Validation

The correlation coefficient values obtained using the Extra trees model are comparable to the values obtained using the RF model, thereby validating the performnace of the RF model.

This implies that with the addition of SCF to the diet, the gut microbiome was affected and this indicates a positive influence on its relationship to bone health.

6.2 Analysis of Microbes

The permutation feature importance technique was used to obtain the importance of each microbe in predicting the six BMD and BMC measures. The microbes were ranked according to their importance values from most important to least important in predicting each of the 6 measures in the two cases of without SCF (0 g/d fiber) and with SCF (12 g/d fiber).

The microbes which had an importance value of above the threshold 0 across all the six bone health measures, were shortlisted to be the most relevant microbes in the two cases. A total 4 microbes were identified as relevant in the case of without SCF treatment and these microbes along with their importance values are shown in Figure 6.1.

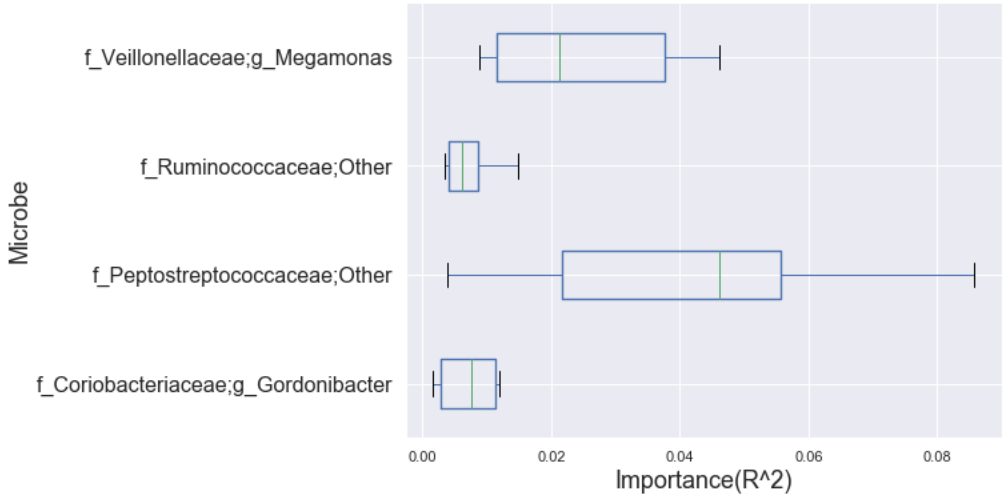


Figure 6.1: 4 Most Relevant Microbes Identified in the CON Treatment (0 g/d Fiber) Across the Six Bone Health Measures.

A total of 22 microbes were identified as relevant in the with SCF treatment case and these microbes along with their importance values are shown in Figure 6.2. It was observed that the microbes *Megamonas (g)* and Unclassified *Peptostreptococcaceae (f)* were common to both the cases. Some of the new microbes added to the list included

Acetivibrio (g), *Prevotella (g)*, *Eubacterium (g)*, and *Turicibacter (g)*.

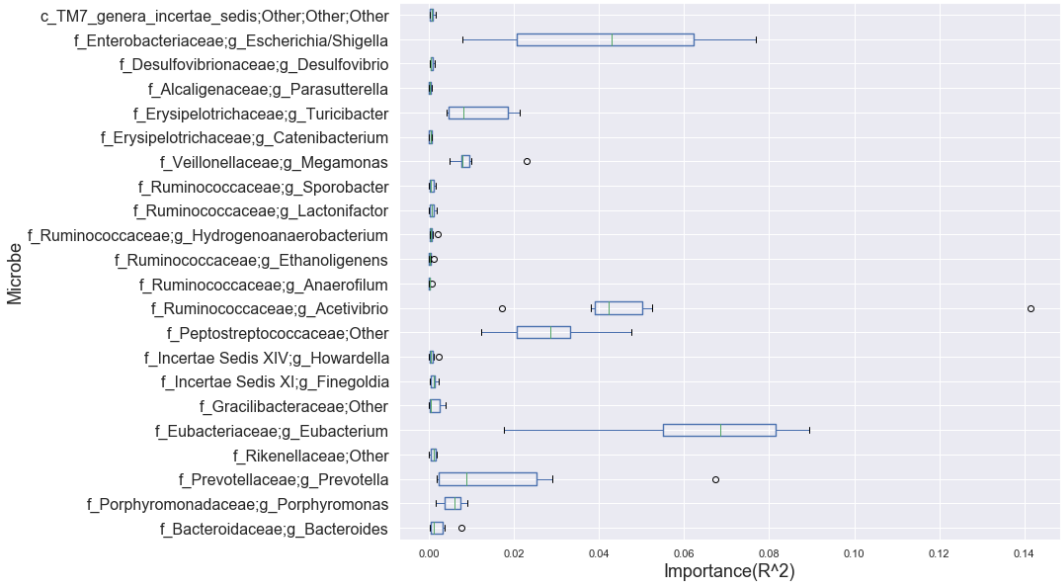


Figure 6.2: 22 Most Relevant Microbes Identified in the SCF Treatment (12 g/d Fiber) Across the Six Bone Health Measures.

The change in importance values of these most relevant microbes with the addition of SCF to the diet is shown in Figure 6.3.

A previous study by Bass *et al.* (1999) in adolescents showed that the due to the difference in speed of growth between different regions of the body there are differences in the size, mass, or BMD in different body parts and there may be a deficiency in one region and not the other. It was seen that growth of the spine accelerated while growth of the legs slowed without a detectable acceleration phase, at puberty.In some other studies it was found that the BMD and BMC measures were significantly higher in the spine region as compared to other body parts (Saraví and Sayegh (2013), Deng *et al.* (2002)).

It was observed from the values of the correlation coefficients obtained in this experiment that, the spine region is of importance and the microbes that play a role in predicting the TSBMD and TSMBC measures were identified.

The top relevant microbes ranked from most important to least important in pre-

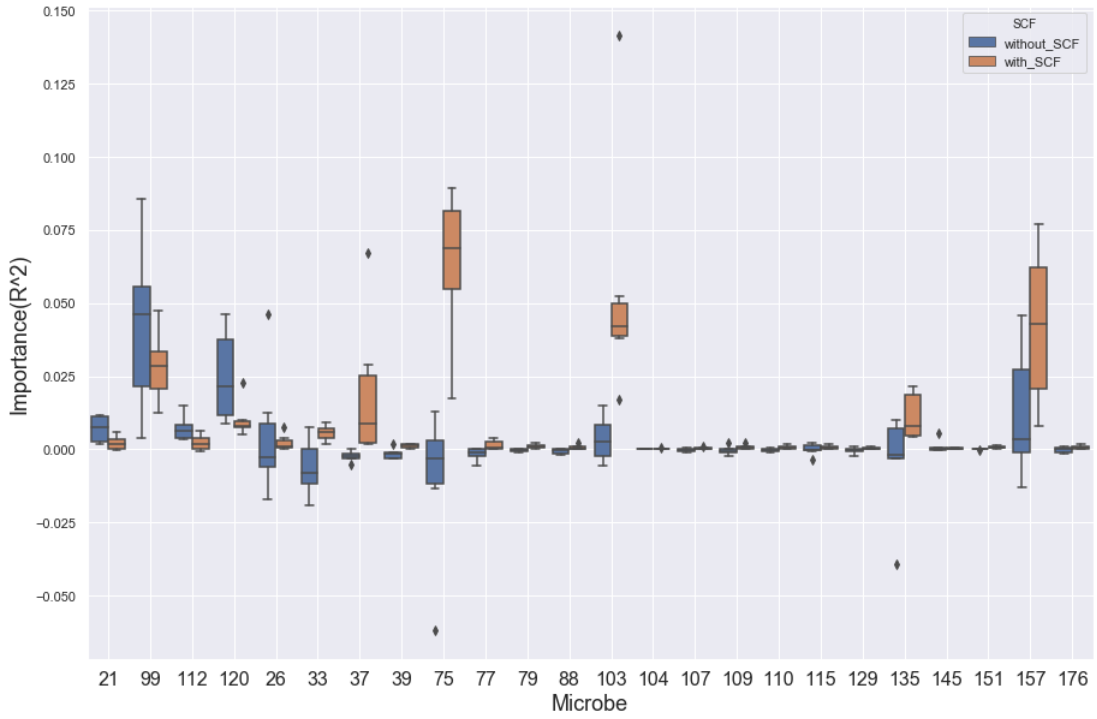


Figure 6.3: The Change in Importance Values With and Without SCF (Microbe Numbers in Table 6.4).

dicting TSBMD and TSBMC measures along with their importance values, in both the cases, are shown in Figure 6.4 and Figure 6.5 respectively.

Some of the most important microbes which were highly ranked in the case of without SCF were, Unclassified *Peptostreptococcaceae (f)*, *Megamonas (g)*, *Phascolarctobacterium (g)*, Unclassified *Ruminococcaceae (f)*, and *Haemophilus (g)*.

The microbes which were ranked high in the case of with SCF were, Unclassified *Firmicutes (p)*, *Acetivibrio (g)*, *Faecalibacterium (g)*, and *Prevotella (g)*, *Eubacterium (g)*, *Turicibacter (g)*, Unclassified *Ruminococcaceae (f)*, *Mogibacterium (g)*, and *Porphyromonas (g)*.

The mean proportions of the most relevant microbes in the faecal samples in both SCF and CON treatments, while the TSBMD and TSBMC measures were recorded, were compared. Figure 6.6 and Figure 6.7 show the mean proportions of the microbes in

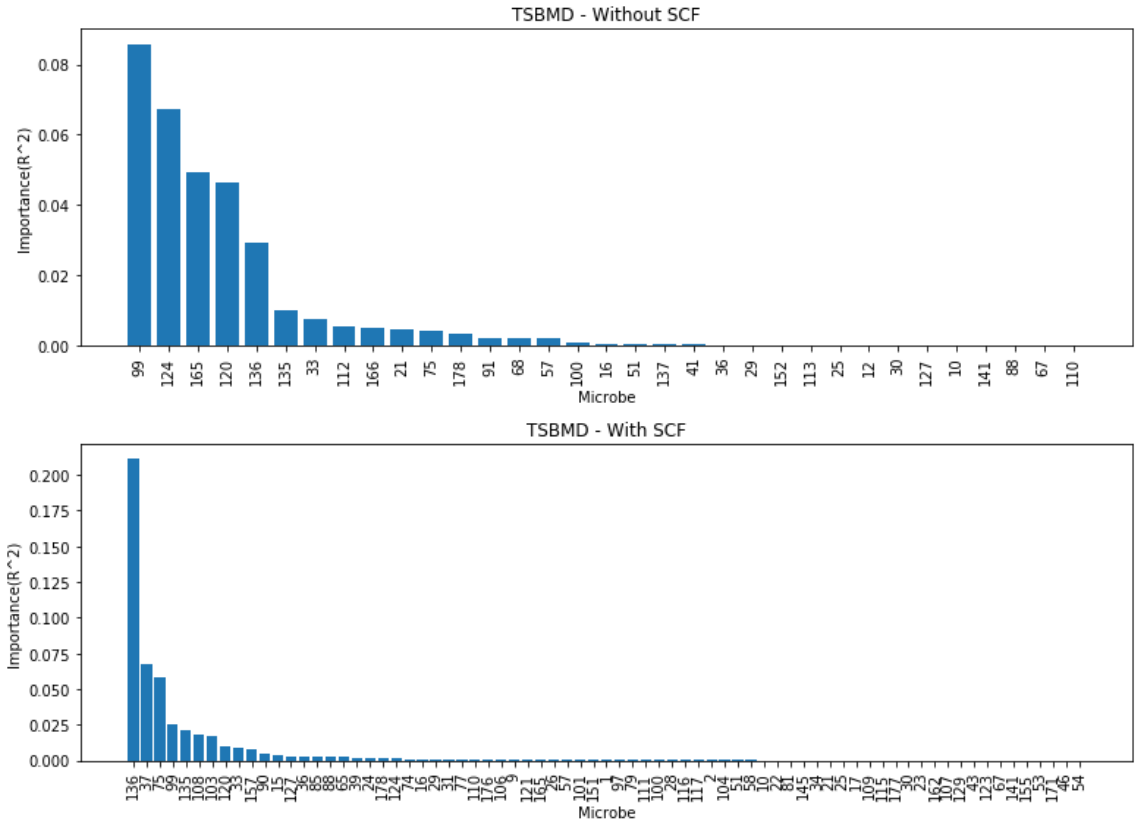


Figure 6.4: Most Relvant Microbes Identified for Predciting TSBMD Measure (Microbe Numbers in Table 6.4).

both the cases of with and without SCF.

Some of the relevant microbes which showed an increase in the mean proportion values with SCF intervention in the diet, while measuring TSBMD, were *Megamonas (g)*, *Faecalibacterium (g)*, *Porphyromonas (g)*, *Anaerostipes (g)*, *Allobaculum*, and *Paraprevotella (g)*. While measuring TSBMC, the microbes *Bacteroides (g)*, Unclassified *Ruminococcaceae (f)*, *Megamonas (g)*, *Faecalibacterium (g)*, *Porphyromonas (g)*, and Unclassified *Firmicutes (p)*, showed an increase in mean proportion values with SCF intervention in the diet.

The most relevant microbes identified in this study, have previously shown to be important for calcium absorption and hence play an important role in affecting the bone health in humans, as discussed in Section 2.5.

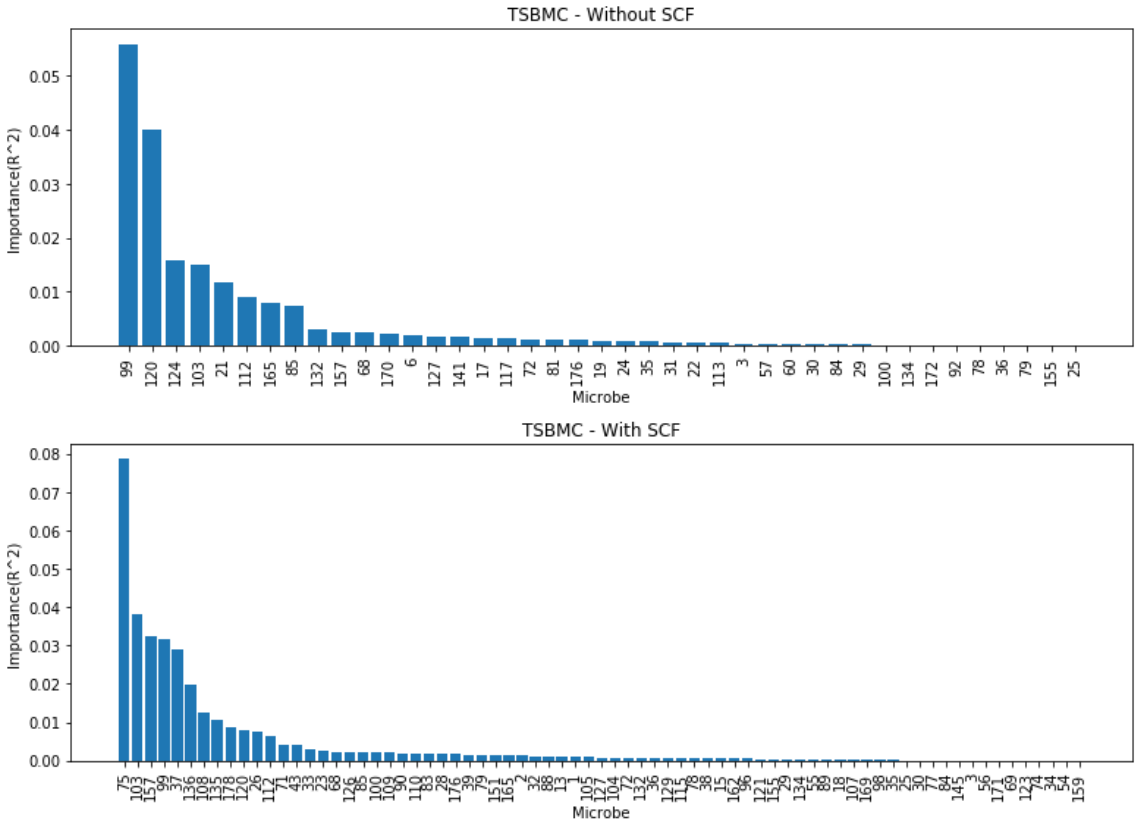


Figure 6.5: Most Relvant Microbes Identified for Predciting TSBMC Measure (Microbe Numbers in Table 6.4).

Microbe Number	Microbe Name
1	*p_Actinobacteria;f_Actinomycetaceae;g_Actinomyces*
2	*p_Actinobacteria;f_Actinomycetaceae;g_Actinomyces*
3	*p_Actinobacteria;f_Actinomycetaceae;Other*
4	*p_Actinobacteria;f_Actinomycetaceae;g_Varibaculum*
5	*p_Actinobacteria;f_Brevibacteriaceae;g_Brevibacterium*
6	*p_Actinobacteria;f_Corynebacteriaceae;g_Corynebacterium*
7	*p_Actinobacteria;f_Corynebacteriaceae;Other*
8	*p_Actinobacteria;f_Micrococcaceae;g_Kocuria*
9	*p_Actinobacteria;f_Micrococcaceae;g_Rothia*

Microbe Number	Microbe Name
10	*p_Actinobacteria;o_Actinomycetales;Other;Other*
11	*p_Actinobacteria;f_Propionibacteriaceae;Other*
12	*p_Actinobacteria;f_Propionibacteriaceae;g_Propionibacterium*
13	*p_Actinobacteria;f_Bifidobacteriaceae;g_Bifidobacterium*
14	*p_Actinobacteria;f_Bifidobacteriaceae;g_Gardnerella*
15	*p_Actinobacteria;f_Bifidobacteriaceae;Other*
16	*p_Actinobacteria;f_Coriobacteriaceae;g_Asaccharobacter*
17	*p_Actinobacteria;f_Coriobacteriaceae;g_Atopobium*
18	*p_Actinobacteria;f_Coriobacteriaceae;g_Collinsella*
19	*p_Actinobacteria;f_Coriobacteriaceae;g_Eggerthella*
20	*p_Actinobacteria;f_Coriobacteriaceae;g_Enterorhabdus*
21	*p_Actinobacteria;f_Coriobacteriaceae;g_Gordonibacter*
22	*p_Actinobacteria;f_Coriobacteriaceae;g_Olsenella*
23	*p_Actinobacteria;f_Coriobacteriaceae;Other*
24	*p_Actinobacteria;f_Coriobacteriaceae;g_Slackia*
25	*p_Actinobacteria;c_Actinobacteria;Other;Other;Other*
26	*p_Bacteroidetes;f_Bacteroidaceae;g_Bacteroides*
27	*p_Bacteroidetes;o_Bacteroidales;Other;Other*
28	*p_Bacteroidetes;f_Porphyromonadaceae;g_Barnesiella*
29	*p_Bacteroidetes;f_Porphyromonadaceae;g_Butyricimonas*
30	*p_Bacteroidetes;f_Porphyromonadaceae;g_Odoribacter*
31	*p_Bacteroidetes;f_Porphyromonadaceae;Other*
32	*p_Bacteroidetes;f_Porphyromonadaceae;g_Parabacteroides*
33	*p_Bacteroidetes;f_Porphyromonadaceae;g_Porphyromonas*

Microbe Number	Microbe Name
34	*p_Bacteroidetes;f_Prevotellaceae;g_Hallella*
35	*p_Bacteroidetes;f_Prevotellaceae;Other*
36	*p_Bacteroidetes;f_Prevotellaceae;g_Paraprevotella*
37	*p_Bacteroidetes;f_Prevotellaceae;g_Prevotella*
38	*p_Bacteroidetes;f_Rikenellaceae;g_Alistipes*
39	*p_Bacteroidetes;f_Rikenellaceae;Other*
40	*p_Bacteroidetes;f_Flavobacteriaceae;g_Cloacibacterium*
41	*p_Bacteroidetes;Other;Other;Other;Other*
42	*p_Bacteroidetes;f_Sphingobacteriaceae;g_Pedobacter*
43	*p_Cyanobacteria;f_Streptophyta;Other*
44	*p_Cyanobacteria;Family I;GpI;Other*
45	*p_Firmicutes;f_Bacillaceae;g_Anoxybacillus*
46	*p_Firmicutes;f_Bacillaceae;g_Bacillus*
47	*p_Firmicutes;f_Bacillaceae;Other*
48	*p_Firmicutes;f_Paenibacillaceae;g_Brevibacillus*
49	*p_Firmicutes;f_Staphylococcaceae;g_Gemella*
50	*p_Firmicutes;f_Staphylococcaceae;g_Macrococcus*
51	*p_Firmicutes;f_Staphylococcaceae;g_Staphylococcus*
52	*p_Firmicutes;f_Thermoactinomycetaceae;g_Desmospora*
53	*p_Firmicutes;f_Aerococcaceae;g_Abiotrophia*
54	*p_Firmicutes;f_Carnobacteriaceae;g_Carnobacterium*
55	*p_Firmicutes;f_Carnobacteriaceae;g_Granulicatella*
56	*p_Firmicutes;f_Carnobacteriaceae;Other*
57	*p_Firmicutes;f_Enterococcaceae;g_Enterococcus*

Microbe Number	Microbe Name
58	*p_Firmicutes;f_Enterococcaceae;Other*
59	*p_Firmicutes;f_Enterococcaceae;g_Vagococcus*
60	*p_Firmicutes;f_Lactobacillaceae;g_Lactobacillus*
61	*p_Firmicutes;f_Lactobacillaceae;Other*
62	*p_Firmicutes;f_Lactobacillaceae;g_Pediococcus*
63	*p_Firmicutes;f_Leuconostocaceae;g_Leuconostoc*
64	*p_Firmicutes;f_Leuconostocaceae;g_Weissella*
65	*p_Firmicutes;o_Lactobacillales;Other;Other*
66	*p_Firmicutes;f_Streptococcaceae;g_Lactococcus*
67	*p_Firmicutes;f_Streptococcaceae;Other*
68	*p_Firmicutes;f_Streptococcaceae;g_Streptococcus*
69	*p_Firmicutes;c_Bacilli;Other;Other;Other*
70	*p_Firmicutes;f_Clostridiaceae;g_Anaerobacter*
71	*p_Firmicutes;f_Clostridiaceae;g_Clostridium*
72	*p_Firmicutes;f_Clostridiaceae;Other*
73	*p_Firmicutes;f_Clostridiaceae;g_Sarcina*
74	*p_Firmicutes;f_Eubacteriaceae;g_Anaerofustis*
75	*p_Firmicutes;f_Eubacteriaceae;g_Eubacterium*
76	*p_Firmicutes;f_Eubacteriaceae;Other*
77	*p_Firmicutes;f_Gracilibacteraceae;Other*
78	*p_Firmicutes;f_Incertae Sedis XI;g_Anaerococcus*
79	*p_Firmicutes;f_Incertae Sedis XI;g_Finegoldia*
80	*p_Firmicutes;f_Incertae Sedis XI;g_Gallicola*
81	*p_Firmicutes;f_Incertae Sedis XI;Other*

Microbe Number	Microbe Name
82	*p_Firmicutes;f_Incertae Sedis XI;g_Parvimonas*
83	*p_Firmicutes;f_Incertae Sedis XI;g_Peptoniphilus*
84	*p_Firmicutes;f_Incertae Sedis XIII;g_Anaerovorax*
85	*p_Firmicutes;f_Incertae Sedis XIII;g_Mogibacterium*
86	*p_Firmicutes;f_Incertae Sedis XIII;Other*
87	*p_Firmicutes;f_Incertae Sedis XIV;g_Blautia*
88	*p_Firmicutes;f_Incertae Sedis XIV;g_Howardella*
89	*p_Firmicutes;f_Incertae Sedis XIV;Other*
90	*p_Firmicutes;f_Lachnospiraceae;g_Anaerostipes*
91	*p_Firmicutes;f_Lachnospiraceae;g_Coprococcus*
92	*p_Firmicutes;f_Lachnospiraceae;g_Dorea*
93	*p_Firmicutes;f_Lachnospiraceae;g_Oribacterium*
94	*p_Firmicutes;f_Lachnospiraceae;Other*
95	*p_Firmicutes;f_Lachnospiraceae;g_Robinsoniella*
96	*p_Firmicutes;f_Lachnospiraceae;g_Roseburia*
97	*p_Firmicutes;o_Clostridiales;Other;Other*
98	*p_Firmicutes;f_Peptococcaceae;g_Peptococcus*
99	*p_Firmicutes;f_Peptostreptococcaceae;Other*
100	*p_Firmicutes;f_Peptostreptococcaceae;g_Peptostreptococcus*
101	*p_Firmicutes;f_Peptostreptococcaceae;g_Sporacetigenium*
102	*p_Firmicutes;f_Ruminococcaceae;g_Acetanaerobacterium*
103	*p_Firmicutes;f_Ruminococcaceae;g_Acetivibrio*
104	*p_Firmicutes;f_Ruminococcaceae;g_Anaerofilum*
105	*p_Firmicutes;f_Ruminococcaceae;g_Anaerotruncus*

Microbe Number	Microbe Name
106	*p_Firmicutes;f_Ruminococcaceae;g_Butyricicoccus*
107	*p_Firmicutes;f_Ruminococcaceae;g_Ethanoligenens*
108	*p_Firmicutes;f_Ruminococcaceae;g_Faecalibacterium*
109	*p_Firmicutes;f_Ruminococcaceae;g_Hydrogenoanaerobacterium*
110	*p_Firmicutes;f_Ruminococcaceae;g_Lactonifactor*
111	*p_Firmicutes;f_Ruminococcaceae;g_Oscillibacter*
112	*p_Firmicutes;f_Ruminococcaceae;Other*
113	*p_Firmicutes;f_Ruminococcaceae;g_Papillibacter*
114	*p_Firmicutes;f_Ruminococcaceae;g_Ruminococcus*
115	*p_Firmicutes;f_Ruminococcaceae;g_Sporobacter*
116	*p_Firmicutes;f_Ruminococcaceae;g_Subdoligranulum*
117	*p_Firmicutes;f_Veillonellaceae;g_Acidaminococcus*
118	*p_Firmicutes;f_Veillonellaceae;g_Allisonella*
119	*p_Firmicutes;f_Veillonellaceae;g_Dialister*
120	*p_Firmicutes;f_Veillonellaceae;g_Megamonas*
121	*p_Firmicutes;f_Veillonellaceae;g_Megasphaera*
122	*p_Firmicutes;f_Veillonellaceae;g_Mitsuokella*
123	*p_Firmicutes;f_Veillonellaceae;Other*
124	*p_Firmicutes;f_Veillonellaceae;g_Phascolarctobacterium*
125	*p_Firmicutes;f_Veillonellaceae;g_Veillonella*
126	*p_Firmicutes;c_Clostridia;Other;Other;Other*
127	*p_Firmicutes;f_Erysipelotrichaceae;g_Allobaculum*
128	*p_Firmicutes;f_Erysipelotrichaceae;g_Bulleidia*
129	*p_Firmicutes;f_Erysipelotrichaceae;g_Catenibacterium*

Microbe Number	Microbe Name
130	*p_Firmicutes;f_Erysipelotrichaceae;g_Coprobacillus*
131	*p_Firmicutes;f_Erysipelotrichaceae;g_Erysipelothrix*
132	*p_Firmicutes;f_Erysipelotrichaceae;g_Holdemania*
133	*p_Firmicutes;f_Erysipelotrichaceae;Other*
134	*p_Firmicutes;f_Erysipelotrichaceae;g_Solobacterium*
135	*p_Firmicutes;f_Erysipelotrichaceae;g_Turicibacter*
136	*p_Firmicutes;Other;Other;Other;Other*
137	*p_Fusobacteria;f_Fusobacteriaceae;g_Fusobacterium*
138	*p_Fusobacteria;f_Fusobacteriaceae;Other*
139	*k_Bacteria;Other;Other;Other;Other;Other*
140	*p_Proteobacteria;f_Caulobacteraceae;g_Brevundimonas*
141	*p_Proteobacteria;Alphaproteobacteria;Other;Other;Other*
142	*p_Proteobacteria;f_Methylocystaceae;g_Methylocystis*
143	*p_Proteobacteria;o_Rhizobiales;Other;Other*
144	*p_Proteobacteria;f_Rhizobiaceae;g_Rhizobium*
145	*p_Proteobacteria;f_Alcaligenaceae;g_Parasutterella*
146	*p_Proteobacteria;f_Alcaligenaceae;g_Sutterella*
147	*p_Proteobacteria;f_Comamonadaceae;g_Diaphorobacter*
148	*p_Proteobacteria;f_Oxalobacteraceae;Other*
149	*p_Proteobacteria;f_Oxalobacteraceae;g_Oxalobacter*
150	*p_Proteobacteria;Neisseriaceae;g_Neisseria*
151	*p_Proteobacteria;f_Desulfovibrionaceae;g_Desulfovibrio*
152	*p_Proteobacteriaf_;Desulfovibrionaceae;Other*
153	*p_Proteobacteria;o_Desulfovibrionales;Other;Other*

Microbe Number	Microbe Name
154	*p_Proteobacteria;o_Alteromonadales;Other;Other*
155	*p_Proteobacteria;f_Shewanellaceae;g_Shewanella*
156	*p_Proteobacteria;f_Enterobacteriaceae;g_Cronobacter*
157	*p_Proteobacteria;f_Enterobacteriaceae;g_Escherichia/Shigella*
158	*p_Proteobacteria;f_Enterobacteriaceae;g_Klebsiella*
159	*p_Proteobacteria;f_Enterobacteriaceae;Other*
160	*p_Proteobacteria;f_Enterobacteriaceae;g_Raoultella*
161	*p_Proteobacteria;f_Enterobacteriaceae;g_Serratia*
162	*p_Proteobacteria;f_Halomonadaceae;g_Halomonas*
163	*p_Proteobacteria;f_Halomonadaceae;Other*
164	*p_Proteobacteria;c_Gammaproteobacteria;Other;Other;Other*
165	*p_Proteobacteria;f_Pasteurellaceae;g_Haemophilus*
166	*p_Proteobacteria;f_Pasteurellaceae;Other*
167	*p_Proteobacteria;f_Moraxellaceae;g_Acinetobacter*
168	*p_Proteobacteria;o_Pseudomonadales;Other;Other*
169	*p_Proteobacteria;f_Pseudomonadaceae;Other*
170	*p_Proteobacteria;f_Pseudomonadaceae;g_Pseudomonas*
171	*p_Proteobacteria;f_Xanthomonadaceae;Other*
172	*p_Proteobacteria;f_Xanthomonadaceae;g_Stenotrophomonas*
173	*p_Proteobacteria;Other;Other;Other;Other*
174	*p_Spirochaetes;f_Spirochaetaceae;g_Treponema*
175	*p_Synergistetes;f_Synergistaceae;g_Cloacibacillus*
176	*p_TM7;c_TM7_genera_incertae_sedis;Other;Other;Other*
177	*p_Verrucomicrobia;f_Verrucomicrobiaceae;g_Akkermansia*

Microbe Number	Microbe Name
178	*None;Other;Other;Other;Other;Other*

Table 6.4: List of Gut Microbes.

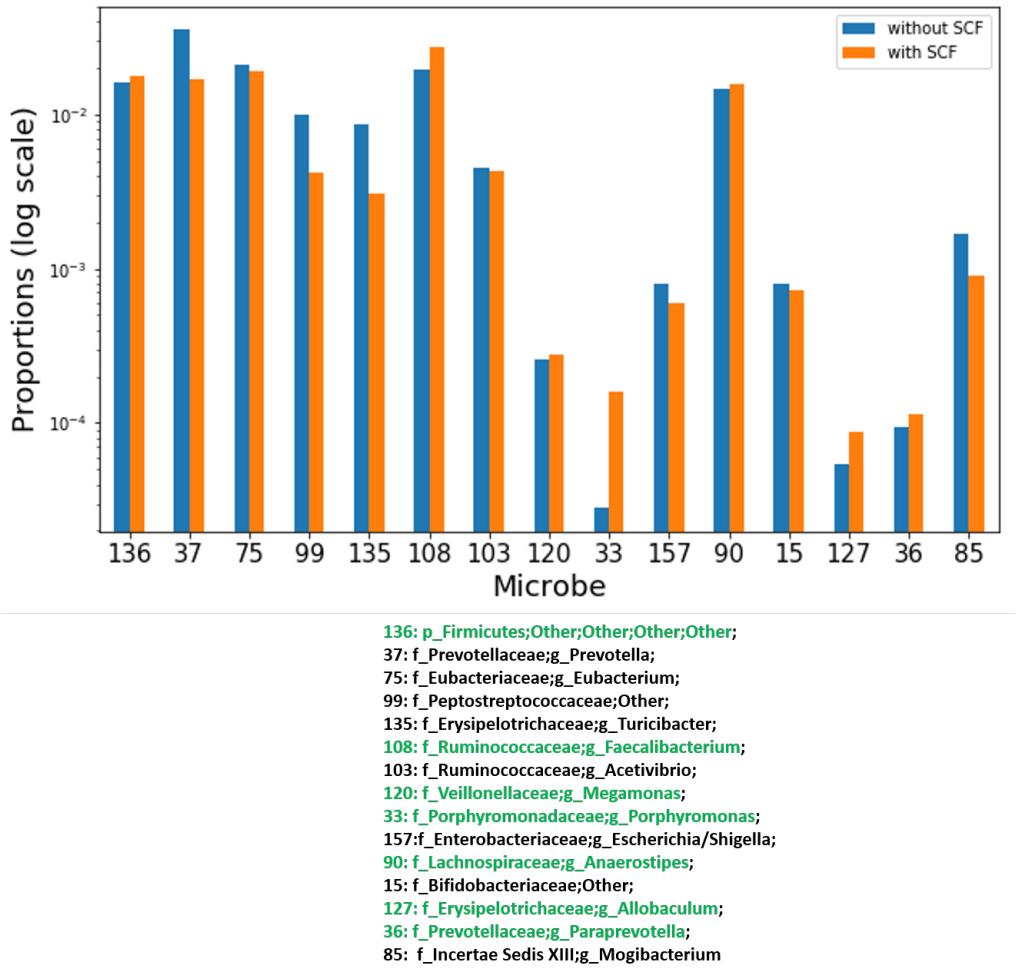
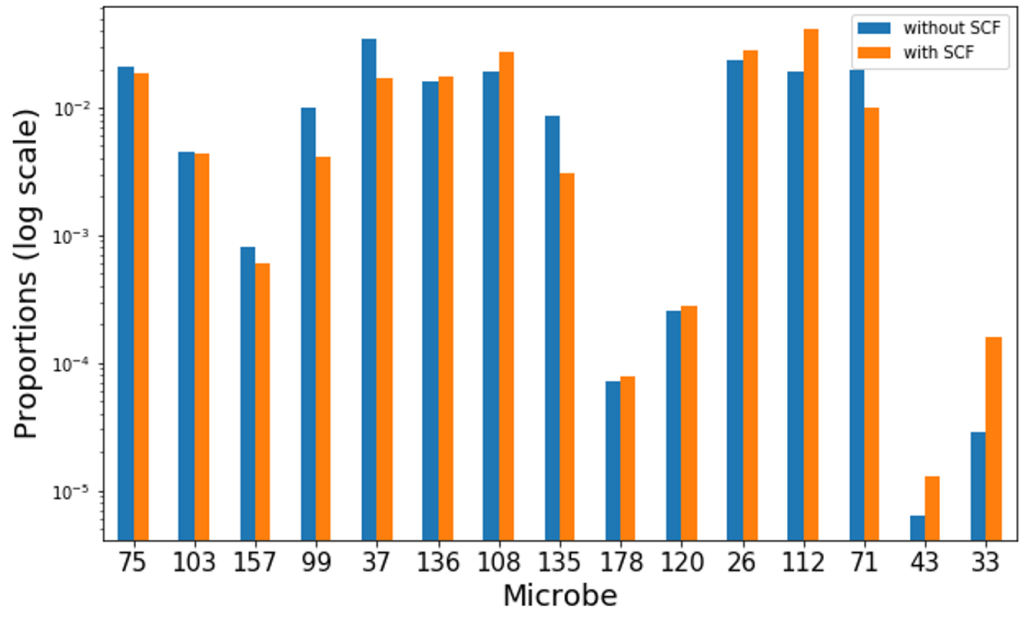


Figure 6.6: Mean Proportions of Most Relevant Microbes for TSBMD. The Microbes in Green Show Increase in Proportion With SCF.



75: f_Eubacteriaceae;g_Eubacterium;
103: f_Ruminococcaceae;g_Acetivibrio;
157:f_Enterobacteriaceae;g_Escherichia/Shigella;
99: f_Peptostreptococcaceae;Other;
37: f_Prevotellaceae;g_Prevotella;
136: p_Firmicutes;Other;Other;Other;Other;
108: f_Ruminococcaceae;g_Faecalibacterium;
135:f_Erysipelotrichaceae;g_Turicibacter;
178: None;Other;Other;Other;Other;Other;
120: f_Veillonellaceae;g_Megamonas;
26: f_Bacteroidaceae;g_Bacteroides;
112: f_Ruminococcaceae;Other;
71: f_Clostridiaceae;g_Clostridium;
43: f_Streptophyta;Other;
33: f_Porphyromonadaceae;g_Porphyromonas

Figure 6.7: Mean Proportions of Most Relevant Microbes for TSBMC. The Microbes in Green Show Increase in Proportion With SCF.

Chapter 7

CONCLUSION

Through this study, a noticeable relationship between the GM community and the overall bone strength was observed. Results showed that there exists a correlation between the gut microbiome content and six measures of BMC and BMD. It was oberved that addition of SCF to the diet increased the correleation between the true and predicted values of BMD and BMC measures ($r \approx 0.51$).

This is one of the few studies where the relationship between GM and bone health has been presented using clinical trials of human participants. It was observed from the values of the correlation coefficients obtained in this experiment that, the spine region is of importance and the microbes that play a role in predicting the TSBMD and TSMBC measures were identified.

It was observed that most of the important microbes belonged to the phyla *Firmicutes* and *Bacteroidetes*. The bacteria of genera *Bacteroides*, *Ruminococcus*, *Prevotella*, *Megamonas*, *Eubacterium*, *Faecalibacterium*, *Bifidobacteria*, *Phascolarctobacterium*, and *Lactobacilli* were some of the most relevant microbes identified in affecting the measures of BMD and BMC. These microbes have shown to be associated with fermentation of prebiotic fibers and play a major role in influencing bone health.

In summary, this study suggests that there exists a relationship between human GM and bone health, and shows that SCF interventions further strengthens the relationship. This study also indicates that this observed relationship is not a causal one.

Future Research

Due to the small sample size of this dataset and short duration of the experiment, significant changes in bone strength and density were not observed. In the future,

studies need to be conducted for longer durations, for one year or more, especially in humans in order to understand the long-term beneficial effects of prebiotic consumption on bone metabolism and bone strength. Collecting samples from a higher number of human subjects will also be useful in these studies to explore the effects on a diverse and larger sample distribution.

It has been shown that consuming SCF is associated with resulting in a healthy microbiome. Further controlled trials are required to investigate the mechanisms by which SCF influences the intestinal microbiota and further affects calcium absorption, bone mineral density, bone mineral content, and bone geometry, over greater lengths of time.

Present research studies show the involvement and importance of changes in the composition of gut microbiota in bone metabolism and improved mineral absorption. Future work is needed to interpret these effects with respect to a variety of prebiotic fibers and their influence with different doses, as well as on various microbial communities.

Therefore, further exploring this relationship will lead to beneficial outcomes in the fields of bone health and osteoporosis research and serve as an important factor for dietary and clinical recommendations.

To further understand the gut-bone relationship, the interactions between the correlated microbes and their combined influence on predicting the bone strength measures need to be addressed by using different feature selection techniques. Different machine learning techniques such as boosting and combining multiple ensemble learning algorithms can further be exploited to gain better insights on this relationship and validate the results obtained.

Abrams, S. A., I. J. Griffin, K. M. Hawthorne, L. Liang, S. K. Gunn, G. Darlington and K. J. Ellis, "A combination of prebiotic short-and long-chain inulin-type fructans enhances calcium absorption and bone mineralization in young adolescents–", The American journal of clinical nutrition **82**, 2, 471–476 (2005).

Ariefdjohan, M. W., D. A. Savaiano and C. H. Nakatsu, "Comparison of dna extraction kits for pcr-dgge analysis of human intestinal microbial communities from fecal specimens", Nutrition journal **9**, 1, 23 (2010).

Bailey, R. L., K. W. Dodd, J. A. Goldman, J. J. Gahche, J. T. Dwyer, A. J. Moshfegh, C. T. Sempos and M. F. Picciano, "Estimation of total usual calcium and vitamin d intakes in the united states", The Journal of nutrition **140**, 4, 817–822 (2010).

Bass, S., P. D. Delmas, G. Pearce, E. Hendrich, A. Tabensky, E. Seeman *et al.*, "The differing tempo of growth in bone size, mass, and density in girls is region-specific", The Journal of clinical investigation **104**, 6, 795–804 (1999).

Bassaganya-Riera, J., M. DiGuardo, M. Viladomiu, A. de Horna, S. Sanchez, A. W. Einerhand, L. Sanders and R. Hontecillas, "Soluble fibers and resistant starch ameliorate disease activity in interleukin-10–deficient mice with inflammatory bowel disease", The Journal of nutrition **141**, 7, 1318–1325 (2011).

Bull, M. J. and N. T. Plummer, "Part 1: The human gut microbiome in health and disease", Integrative Medicine: A Clinician's Journal **13**, 6, 17 (2014).

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon *et al.*, "Qiime allows analysis of high-throughput community sequencing data", Nature methods **7**, 5, 335–336 (2010).

Carey, D. E. and N. H. Golden, "Bone health in adolescence.", Adolescent medicine: state of the art reviews **26**, 2, 291 (2015).

Chen, Y.-C., J. Greenbaum, H. Shen and H.-W. Deng, "Association between gut microbiota and bone health: potential mechanisms and prospective", The Journal of Clinical Endocrinology & Metabolism **102**, 10, 3635–3646 (2017).

Cheng, S., X. Qi, M. Ma, L. Zhang, B. Cheng, C. Liang, L. Liu, P. Li, O. P. Kafle, Y. Wen *et al.*, "Assessing the relationship between gut microbiota and bone mineral density", Frontiers in genetics **11**, 6 (2020).

Clarke, G., R. M. Stilling, P. J. Kennedy, C. Stanton, J. F. Cryan and T. G. Dinan, "Minireview: gut microbiota: the neglected endocrine organ", Molecular endocrinology **28**, 8, 1221–1238 (2014).

David, L. A., C. F. Maurice, R. N. Carmody, D. B. Gootenberg, J. E. Button, B. E. Wolfe, A. V. Ling, A. S. Devlin, Y. Varma, M. A. Fischbach *et al.*, "Diet rapidly and reproducibly alters the human gut microbiome", Nature **505**, 7484, 559–563 (2014).

Deng, H.-W., F.-H. Xu, K. M. Davies, R. Heaney and R. R. Recker, "Differences in bone mineral density, bone mineral content, and bone areal size in fracturing and non-fracturing women, and their interrelationships at the spine and hip", Journal of bone and mineral metabolism **20**, 6, 358–366 (2002).

Devareddy, L., D. A. Khalil, K. Korlagunta, S. Hooshmand, D. D. Bellmer and B. H. Arjmandi, "The effects of fructo-oligosaccharides in combination with soy protein on bone in osteopenic ovariectomized rats", Menopause **13**, 4, 692–699 (2006).

Fukui, H., A. Nishida, S. Matsuda, F. Kira, S. Watanabe, M. Kuriyama, K. Kawakami, Y. Aikawa, N. Oda, K. Arai *et al.*, "Usefulness of machine learning-based gut microbiome analysis for identifying patients with irritable bowels syndrome", Journal of clinical medicine **9**, 8, 2403 (2020).

Fulgoni III, V. L., D. R. Keast, N. Auestad and E. E. Quann, "Nutrients from dairy foods are difficult to replace in diets of americans: food pattern modeling and an analyses of the national health and nutrition examination survey 2003-2006", Nutrition research **31**, 10, 759–765 (2011).

Griffin, I., P. Davila and S. Abrams, "Non-digestible oligosaccharides and calcium absorption in girls with adequate calcium intakes", British Journal of Nutrition **87**, S2, S187–S191 (2002).

Guarner, F. and J.-R. Malagelada, "Gut flora in health and disease", The Lancet **361**, 9356, 512–519 (2003).

Gunduz, N. and E. Fokoué, "Robust classification of high dimension low sample size data", arXiv preprint arXiv:1501.00592 (2015).

Guo, Y., A. Graber, R. N. McBurney and R. Balasubramanian, "Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms", BMC bioinformatics **11**, 1, 447 (2010).

Hastie, T., R. Tibshirani and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction* (Springer Science & Business Media, 2009).

James, G., D. Witten, T. Hastie and R. Tibshirani, *An introduction to statistical learning*, vol. 112 (Springer, 2013).

Johnson, H. R., D. D. Trinidad, S. Guzman, Z. Khan, J. V. Parziale, J. M. DeBruyn and N. H. Lents, "A machine learning approach for using the postmortem skin microbiome to estimate the postmortem interval", PloS one **11**, 12, e0167370 (2016).

Klosterbuer, A. S., M. A. Hullar, F. Li, E. Traylor, J. W. Lampe, W. Thomas and J. L. Slavin, "Gastrointestinal effects of resistant starch, soluble maize fibre and pullulan in healthy adults", British journal of nutrition **110**, 6, 1068–1074 (2013).

Knapp, B. K., L. L. Bauer, K. S. Swanson, K. A. Tappenden, G. C. Fahey and M. R. De Godoy, "Soluble fiber dextrin and soluble corn fiber supplementation modify indices of health in cecum and colon of sprague-dawley rats", Nutrients **5**, 2, 396–410 (2013).

Krebs-Smith, S. M., P. M. Guenther, A. F. Subar, S. I. Kirkpatrick and K. W. Dodd, "Americans do not meet federal dietary recommendations", The Journal of nutrition **140**, 10, 1832–1838 (2010).

Le Chatelier, E., T. Nielsen, J. Qin, E. Prifti, F. Hildebrand, G. Falony, M. Almeida, M. Arumugam, J.-M. Batto, S. Kennedy *et al.*, "Richness of human gut microbiome correlates with metabolic markers", Nature **500**, 7464, 541–546 (2013).

Levine, M. A., "Assessing bone health in children and adolescents", Indian journal of endocrinology and metabolism **16**, Suppl 2, S205 (2012).

Lobo, A. R., C. Colli and T. M. Filisetti, "Fructooligosaccharides improve bone mass and biomechanical properties in rats", Nutrition research **26**, 8, 413–420 (2006).

Loud, K. J. and C. M. Gordon, "Adolescent bone health", Archives of pediatrics & adolescent medicine **160**, 10, 1026–1032 (2006).

Louppe, G., "Understanding random forests: From theory to practice", arXiv preprint arXiv:1407.7502 (2014).

Luan, J., C. Zhang, B. Xu, Y. Xue and Y. Ren, "The predictive performances of random forest models with limited sample size and different species traits", Fisheries Research **227**, 105534 (2020).

Maathuis, A., A. Hoffman, A. Evans, L. Sanders and K. Venema, "The effect of the undigested fraction of maize products on the activity and composition of the microbiota determined in a dynamic in vitro model of the human proximal large intestine", Journal of the American College of Nutrition **28**, 6, 657–666 (2009).

Macfarlane, G. T. and S. Macfarlane, "Bacteria, colonic fermentation, and gastrointestinal health", Journal of AOAC International **95**, 1, 50–60 (2012).

Martin, F.-P. J., N. Sprenger, I. Montoliu, S. Rezzi, S. Kochhar and J. K. Nicholson, "Dietary modulation of gut functional ecology studied by fecal metabonomics", Journal of proteome research **9**, 10, 5284–5295 (2010).

McCabe, L., R. A. Britton and N. Parameswaran, "Prebiotic and probiotic regulation of bone health: role of the intestine and its microbiome", Current osteoporosis reports **13**, 6, 363–371 (2015).

Medina-Gomez, C., "Bone and the gut microbiome: a new dimension", J Lab Precis Med **3**, 96 (2018).

Meding, S., U. Nitsche, B. Balluff, M. Elsner, S. Rauser, C. Schone, M. Nipp, M. Maak, M. Feith, M. P. Ebert *et al.*, "Tumor classification of six common cancer types based on proteomic profiling by maldi imaging", Journal of proteome research **11**, 3, 1996–2003 (2012).

Morgan, X. C., N. Segata and C. Huttenhower, "Biodiversity and functional genomics in the human microbiome", Trends in genetics **29**, 1, 51–58 (2013).

Namkung, J., "Machine learning methods for microbiome studies", Journal of Microbiology **58**, 3, 206–216 (2020).

Neish, A. S., "Microbes in gastrointestinal health and disease", Gastroenterology **136**, 1, 65–80 (2009).

Nossa, C. W., W. E. Oberdorf, L. Yang, J. A. Aas, B. J. Paster, T. Z. DeSantis, E. L. Brodie, D. Malamud, M. A. Poles and Z. Pei, "Design of 16s rrna gene primers for 454 pyrosequencing of the human foregut microbiome", World journal of gastroenterology: WJG **16**, 33, 4135 (2010).

Ohlsson, C. and K. Sjögren, "Effects of the gut microbiota on bone mass", Trends in Endocrinology & Metabolism **26**, 2, 69–74 (2015).

Probst, P., M. N. Wright and A.-L. Boulesteix, "Hyperparameters and tuning strategies for random forest", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **9**, 3, e1301 (2019).

Qin, J., Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen *et al.*, "A metagenome-wide association study of gut microbiota in type 2 diabetes", Nature **490**, 7418, 55–60 (2012).

Quigley, E. M., "Gut bacteria in health and disease", Gastroenterology & hepatology **9**, 9, 560 (2013).

Raschka, S., "Model evaluation, model selection, and algorithm selection in machine learning", (2018).

Saraví, F. D. and F. Sayegh, "Bone mineral density and body composition of adult premenopausal women with three levels of physical activity", Journal of osteoporosis **2013** (2013).

Scholz-Ahrens, K. E., Y. Açil and J. Schrezenmeir, "Effect of oligofructose or dietary calcium on repeated calcium and phosphorus balances, bone mineralization and trabecular structure in ovariectomized rats", British Journal of Nutrition **88**, 4, 365–377 (2002).

Sears, C. L., "A dynamic partnership: celebrating our gut flora", Anaerobe **11**, 5, 247–251 (2005).

Shen, S. and C. H. Wong, "Bugging inflammation: role of the gut microbiota", Clinical & translational immunology **5**, 4, e72 (2016).

Sjögren, K., C. Engdahl, P. Henning, U. H. Lerner, V. Tremaroli, M. K. Lagerquist, F. Bäckhed and C. Ohlsson, "The gut microbiota regulates bone mass in mice", Journal of bone and mineral research **27**, 6, 1357–1367 (2012).

Topçuoğlu, B. D., N. A. Lesniak, M. T. Ruffin, J. Wiens and P. D. Schloss, "A framework for effective application of machine learning to microbiome-based classification problems", Mbio **11**, 3 (2020).

Turnbaugh, P. J., V. K. Ridaura, J. J. Faith, F. E. Rey, R. Knight and J. I. Gordon, "The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice", Science translational medicine **1**, 6, 6ra14–6ra14 (2009).

Ursell, L. K., J. L. Metcalf, L. W. Parfrey and R. Knight, "Defining the human microbiome", Nutrition reviews **70**, suppl_1, S38–S44 (2012).

van den Heuvel, E. G., M. H. Schoterman and T. Muijs, "Transgalactooligosaccharides stimulate calcium absorption in postmenopausal women", The Journal of nutrition **130**, 12, 2938–2942 (2000).

Wallace, T. C., M. Marzorati, L. Spence, C. M. Weaver and P. S. Williamson, "New frontiers in fibers: innovative and emerging research on the gut microbiome and bone health", Journal of the American College of Nutrition **36**, 3, 218–222 (2017).

Wang, H., I.-S. Lee, C. Braun and P. Enck, "Effect of probiotics on central nervous system functions in animals and humans: a systematic review", Journal of neurogastroenterology and motility **22**, 4, 589 (2016).

Wang, J., Y. Wang, W. Gao, B. Wang, H. Zhao, Y. Zeng, Y. Ji and D. Hao, "Diversity analysis of gut microbiota in osteoporosis and osteopenia patients", PeerJ **5**, e3450 (2017).

Weaver, C. M., "Diet, gut microbiome, and bone health", Current osteoporosis reports **13**, 2, 125–130 (2015).

Weaver, C. M., B. R. Martin, C. H. Nakatsu, A. P. Armstrong, A. Clavijo, L. D. McCabe, G. P. McCabe, S. Duignan, M. H. Schoterman and E. G. van den Heuvel, "Galactooligosaccharides improve mineral absorption and bone properties in growing rats through gut fermentation", Journal of agricultural and food chemistry **59**, 12, 6501–6510 (2011).

Weaver, C. M., B. R. Martin, J. A. Story, I. Hutchinson and L. Sanders, "Novel fibers increase bone calcium content and strength beyond efficiency of large intestine fermentation", Journal of agricultural and food chemistry **58**, 16, 8952–8957 (2010).

Whisner, C. M. and L. F. Castillo, "Prebiotics, bone and mineral metabolism", Calcified Tissue International **102**, 4, 443–479 (2018).

Whisner, C. M., B. R. Martin, C. H. Nakatsu, G. P. McCabe, L. D. McCabe, M. Peacock and C. M. Weaver, "Soluble maize fibre affects short-term calcium absorption in adolescent boys and girls: a randomised controlled trial using dual stable isotopic tracers", British Journal of Nutrition **112**, 3, 446–456 (2014).

Whisner, C. M., B. R. Martin, C. H. Nakatsu, J. A. Story, C. J. MacDonald-Clarke, L. D. McCabe, G. P. McCabe and C. M. Weaver, "Soluble corn fiber increases calcium absorption associated with shifts in the gut microbiome: a randomized dose-response trial in free-living pubertal females", The Journal of nutrition **146**, 7, 1298–1306 (2016).

Yang, I., E. J. Corwin, P. A. Brennan, S. Jordan, J. R. Murphy and A. Dunlop, "The infant microbiome: implications for infant health and neurocognitive development", Nursing research **65**, 1, 76 (2016).

Yazdani, M., B. C. Taylor, J. W. Debelius, W. Li, R. Knight and L. Smarr, "Using machine learning to identify major shifts in human gut microbiome protein family abundance in disease", in "2016 IEEE international conference on big data (big data)", pp. 1272–1280 (IEEE, 2016).

Zafar, T. A., C. M. Weaver, Y. Zhao, B. R. Martin and M. E. Wastney, "Nondigestible oligosaccharides increase calcium absorption and suppress bone resorption in ovariectomized rats", The Journal of nutrition **134**, 2, 399–402 (2004).

Zeller, G., J. Tap, A. Y. Voigt, S. Sunagawa, J. R. Kultima, P. I. Costea, A. Amiot, J. Böhm, F. Brunetti, N. Habermann *et al.*, "Potential of fecal microbiota for early-stage detection of colorectal cancer", Molecular systems biology **10**, 11, 766 (2014).

Zheng, A., "Evaluating machine learning models: a beginner's guide to key concepts and pitfalls", (2015).

Zhou, Y.-H. and P. Gallins, "A review and tutorial of machine learning methods for microbiome host trait prediction", Frontiers in Genetics **10**, 579 (2019).